Advanced Engineering Informatics

A case study using PCA and signal processing techniques for processing time-based construction settlement data of road embankments

--Manuscript Draft--

Manuscript Number:	ADVEI-D-20-00344R1
Article Type:	Full Length Article
Keywords:	Embankment Construction; Soil Settlement; Data Pre-processing; Principal Component Analysis; Signal Processing; Data Cleaning
Corresponding Author:	Faisal Siddiqui Teesside University Middlesbrough, UNITED KINGDOM
First Author:	Faisal Siddiqui
Order of Authors:	Faisal Siddiqui
	Paul Sargent
	Gary Montague
Abstract:	Instrumentation is beneficial in civil engineering for monitoring structures during their construction and operation. The data collected can be used to observe real-time response and develop data-driven models for predicting future behaviour. However, a limited number of sensors are usually used for on-site civil engineering construction due to cost restrictions and practicalities. This results in relatively small raw datasets, which often contain errors and anomalies. Interpreting and making judicious use of the available dataset for developing reliable predictive model represents a significant challenge. Therefore, it is essential to pre-process and clean the data for improving their quality. To date, little investigation has been performed in the application of such data cleaning methods to geotechnical engineering datasets collected from full-scale sites. The purpose of this study is to apply simple and effective data pre-processing techniques to site-data collected from a highway embankment constructed on a sequence of soil layers of different physical make-up and non-linear consolidation characteristics. Various cleaning methods were applied to magnetic extensometer data collected for monitoring settlement within foundation soils beneath the embankment. PCA was used to explore raw data, identify and remove outliers. Numerous filtering and smoothing methods were used to clean noise in the data and their results were further compared using RMSE and NMSE. The methods adopted for data pre-processing and cleaning proved very effective for capturing the raw settlement behaviour on site. The findings from this study would be useful to site engineers regarding complex decision-making relating to ground response due to embankment construction. This also has positive prospects for developing dynamic prediction models for embankment settlement.
Response to Reviewers:	

Cover Letter

Faisal Siddiqui
PhD Researcher in Civil Engineering

School of Computing, Engineering & Digital Technologies
Teesside University
Middlesbrough
Tees Valley
TS1 3BX, UK

E: f.siddiqui@tees.ac.uk

2nd May 2020

Dear Prof. CH Chen, Prof. T Hartmann and colleagues,

It is our intention to publish our most recent work in your Advanced Engineering Informatics journal. The paper entitled "A case study using PCA and signal processing techniques for processing time-based construction settlement data of road embankments", forms a component of the PhD research being undertaken by the first author.

In this paper, we present a framework for preprocessing and cleaning geotechnical data collected during the construction of a highway embankment in north east England, UK. PCA and smoothing methods were successful in removing outliers and noise from the data and in identifying trends, in preparation for predictive modelling. We believe that this research is appropriate for publication in your journal as it focusses on the application of computational methods for gaining insights into the behaviour of geotechnical structures. The paper should be of interest to your readers who specialise in the areas of geotechnics, data science and data-driven monitoring of civil infrastructures.

We would like to confirm that the research presented in our paper has not been published elsewhere and that we would be very appreciative if it were to be published in your next available edition.

We look forward to hearing from you soon.

Yours sincerely,

Mr Faisal Siddiqui Dr Paul Sargent Prof Gary Montague

Siddiqui et al. (2020): A case study using PCA and signal processing techniques for processing time-based construction settlement data of road embankments.

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
Editor			
1	As this is only a short summary of the most important comments it is self-speaking to address all of the reviewer comments in a detailed manner by thoroughly and globally revising the manuscript and its structure. To allow us to more easily track how you addressed the review comments, I also would like to ask you to submit a table summarizing how you addressed the reviewers' comments in the following format: a column of the verbatim reviewer comment and a column with a detailed description of how you addressed the comment (indicating the page and line number in the revised manuscript where the comment was addressed so that the reviewers and I can easily locate the change)	Many thanks to the editor for this comment. The authors have produced this document, which is a tabulated summary of all reviewer's comments, authors responses and any amendments made to the manuscript (including page number and line number references where appropriate).	N/A
2	Interpreting the reviewer comments, the first reviewer has some issues with the technical aspects of the paper that should be thoroughly addressed in the revision.	Please see the authors' responses to all of Reviewer 1's comments below, which have been addressed in full and	N/A

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		hopefully will add to the clarity and technical content of our manuscript.	
3	Looking at the comments, I also think that the character of the case study approach needs to be made much more explicit throughout the paper in terms of critically reflecting on the generality of the applied methods and in terms of integrating the work stringer into the existing literature.	Many thanks to the Editor for raising this comment. The authors believe that Reviewer 1 has also raised similar comments. In response to this, the authors have inserted new text which provides further reflection on outlier detection and the performance of filtering methods used on the settlement-time data presented in this study. Please see our detailed responses to Reviewer 1 below.	N/A
4	Important for the scope of the journal is that the paper can provide clear evidence that the computational method can support complex engineering tasks.	The authors would like to thank the Editor for this comment. The purpose of this study is applying commercially available data pre-processing techniques (smoothing / filtering) to site-collected data from a complex engineering system (in this case, being a road embankment constructed on a series of soil layers of	Abstract updated

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		different physical make-up and possess non-linear consolidation characteristics). The engineering challenge at hand focusses on interpreting data collected from a very limited number of sensors, which are not state-of-the-art digital in nature due to cost implications and practicalities associated with on-site civil engineering construction. Based on the results presented in our manuscript, the authors believe that they successfully demonstrate that the computational method used was appropriate for capturing the settlement behaviour of the road embankment case study.	
5	Data in Brief (optional): We invite you to convert your supplementary data (or a part of it) into an additional journal publication in Data in Brief, a multi-disciplinary open access journal. Data in Brief articles are a fantastic way to describe	Unfortunately, the raw data is not owned by the authors but by Northumberland County Council and AECOM Environment and Ground Engineering. Therefore, they cannot be made available as open data. If someone wants	N/A

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
	supplementary data and associated metadata, or full raw datasets deposited in an external repository, which are otherwise unnoticed. A Data in Brief article (which will be reviewed, formatted, indexed, and given a DOI) will make your data easier to find, reproduce, and cite.	the data, they can contact the corresponding author. The authors can correspond with AECOM and request for the data. Please note that their availability is at the discretion of the data owners.	
Reviewer 1			
A	This paper investigates applying various cleaning methods on magnetic extensometer data collected for monitoring changes in settlement within foundation soils beneath a highway embankment constructed in Northumberland, UK. From the point of view of methodology, principal component analysis and various filtering methods are common and have not been improved. From the result demonstration, the validity of data preprocessing has not been fully proved. Most importantly, the method is not universal and the author does not suggest its use. In general,	The authors would like to thank the reviewer for their very thorough review of our submission. To clarify, our paper aims to utilise commonly used data driven methods in improving the information gained from site-based geotechnical data. In the field of geotechnical engineering, application of data analysis is not a common approach for studying the consolidation behaviour of embankment structures. This paper demonstrates that the commonly used instrumentation types	Page 7, Line no. 108-113 "*The paper presents a methodology for pre-processing in preparing geotechnical monitoring instrumentation data for a highway embankment case study, with a view to developing a data-driven model and investigate statistical methods to enhance quality and make better settlement predictions. Commonly used PCA, filtering and smoothing methods were adopted for processing the data to encourage their use as an*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
	this paper is not innovative enough to be published in this high-level journal.	and methods of data analysis can be used to extract information. Whilst the reviewer is correct in that PCA is a well-established concept, its use in analysing geotechnical data is currently in its infancy. Please note that the PCA and filtering methods adopted in this study are indeed universal and applicable for analysing data collected from other embankments. However, the values that have been used for 'tuning' the filter parameters applied to the datasets are not universal but can be obtained through judicious application to the data. Therefore, on the basis that PCA and the filtering/smoothing techniques are not commonly used approaches for analysing embankment settlement data	*industrial standard for processing raw geotechnical data."* Page 7, Line no. 115-116 *"Although the concepts of data pre-processing and cleaning used are well established in other domains, they are still in their infancy in the field of geotechnical engineering."*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		and that we have demonstrated their use on industrial data, we believe that this work provides an innovative contribution to the field of study and therefore worthy of publication.	
1	In the introduction, the author uses a case to illustrate that the results of finite element analysis are not reliable, which is not appropriate. In addition, embedding monitoring equipment is also expensive compared with laboratory testing. Therefore, there are some arguments in the article that I cannot agree with.	Many thanks for raising this comment. To clarify, the authors are not suggesting that results from FEA are unreliable. FEA serves as a very valuable tool in engineering design. In our manuscript, we have simply presented a brief overview of the results obtained by Kelly et al. (2018) on the comparison between FEA predicted and site-observed settlement for the Ballina test embankment. There are many studies where site monitoring data have been used to compare against results from FEA results, and there are generally discrepancies between them. However,	Page 5, Line no. 60-66 *"This demonstrates that whilst FEA is a powerful tool in design, it does not consistently provide accurate predictions of site behaviour. For validation, they require accurate site information and soil parameters (e.g. small strain stiffness from triaxial testing) that can be derived from expensive laboratory testing techniques [5]. An alternative method is to calibrate FEA models by updating soil parameters using statistical methods to improve prediction capability. This requires on-site measurement data*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		the level of discrepancy observed is often impacted by a combination of the underpinning assumptions made for certain model parameters and the constitutive model used for predicting soil behaviour. In addition, FEA alone is not sufficiently accurate for on-site decision making during the construction phase. There are a few studies (Zheng et al., 2018; Muthing et al., 2018; Kelly and Huang, 2015) where monitoring data has been used to update FE models during construction, so that their prediction results can be corrected. Certain geotechnical FEA models use advanced constitutive soil models (e.g. modified Cam Clay, Hardening Soil with Small Strain), which require the	*through the installation of ME and VWP instruments [8–10].”*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		input of parameters that cannot be obtained from conventional site-based instrumentation – only from advanced field tests (e.g. pressure meter) or laboratory tests (e.g. hollow cylinder, triaxial with small strain measurements). These advanced tests are indeed very expensive (due to the levels of expertise and time required to complete tests) in comparison with the cost of installing conventional monitoring equipment on site. These previous FEA-based studies assist in forming the basis of this paper, whereby the site data and PCA/data smoothing-filtering approaches adopted attempt to complement FEA findings.	
2	There is a lot of conceptual content in this paper, including the introduction of monitoring instruments and filtering methods. Some	The authors would like to thank the reviewer's comment. We have inserted additional text at the end of the	Page 7-8, Line no. 117-128 "*The structure of the paper is organised as follows. Section 2 presents the case*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
	contents can be cut. In addition, the structure of this paper is chaotic, which makes it difficult for me to read. I suggest modifying the paper structure.	introduction section (Lines 117-128) that aims to briefly outline the structure of the paper. In addition, we acknowledge that there are some sections of text which are not needed and have been removed. Sections 2 and 3 have also been swapped around, whereby details of the case study are presented first in the manuscript, before theoretical background.	*study used for this paper, which includes a description of the instrumentation used, their location and data acquisition frequency. Section 3 provides a theoretical background on data pre-processing in preparation for data analysis, specifically focusing on PCA, signal processing-based filtering and smoothing that have been utilised in this paper for exploring and processing raw data. Section 4 presents the methodology used for investigating and improving the quality of raw embankment settlement data, including: (i) PCA for initial investigation and detecting outliers; (ii) filtering and smoothing to remove noise from the data and improve their quality. Section 5 discusses the data pre-processing results and their statistical comparison using RMSE and NMSE. Section 6 discusses*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
			the overall result and their credibility. Section 7 provides a summary of key findings from the paper and recommendations for future processing of raw geotechnical data." Page 8, Line 129. Section 2: *Case study – Morpeth Northern Bypass, Northumberland, UK* Page 12, Line 171. Section 2.1 *Instrumentation (Magnetic Extensometer and Vibrating wire piezometer as its subsections)* Page 14, Line 225. Section 3: *Theoretical Background* Page 15 Paragraph deleted – after Line no. 235 and after 239.

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
3	How are the parameters of each filter iteratively adjusted, and the corresponding technical implementation is not mentioned in this paper? The author is suggested to add this part of content.	Many thanks for highlighting this issue. The reviewer highlights a challenging issue regarding filter parameter selection. To address this, the authors have inserted some additional text to the manuscript.	Page 24-25, Line no. 401-404 *"The parameters used for each filter were iteratively adjusted based on the geological knowledge and observations made by the engineering expert monitoring the settlement data. The adjustments were made to capture the behaviour of the soil whilst removing any associated noise from the instruments."*
4	Why are two thresholds (AEC and CAEC) set in Figure 3? Which line shall prevail, red or green? In my opinion, either one is well, and the two thresholds are a bit redundant.	Thank you for raising this comment. We have presented two thresholds in Figure 3 to serve as a comparison – whereby the AEC threshold is more conventionally used. Whereas the CAEC threshold is considered to be more conservative (as stated in Lines 329-331 & 350-355). Furthermore, we believe that using these two thresholds provides a higher level of sophistication	N/A

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		compared with using just one threshold. Nevertheless, these metrics alone are insufficient, given that we are searching for subtle behaviour in ground settlement. Hence, we have also used RMSECV in Figure 3. The authors believe that we have already explained the above in the manuscript, and therefore do not see the need to further amend the document.	
5	What is the impact of data cluster overlap? How does the author deal with it? In addition, after eliminating outliers, is linear interpolation the best filling effect?	Thank you for raising this comment. The authors would like to forward the reviewer to Lines 364-368, where we explain that for the data overlap can be expected due to the incremental settlement reducing as the embankment is constructed, as a result of progressive foundation soil consolidation.	Page 22, Line no. 374-378 *"Settlement values of these time data points were removed from the dataset and approximated by linear interpolation. This was considered appropriate since the data sampling rate was relatively high compared with the dynamics of the embankment settlement. Thus, over a short time period compared*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		To find missing values between the available data points, we believe that linear interpolation is appropriate as the data sampling rate was relatively high compared with the dynamics of the embankment settlement. Therefore, if only a few data points were missing, then the short-term settlement experienced by the embankment is relatively small. Thus, linear interpolation would provide a reasonable approximation. To reflect this, the authors have inserted some additional text.	*to the dynamic response, the error associated with linear interpolation will be small and negates the need for higher order interpolation methods."*
6	The paper says that the outliers outside 95% confidence limit are the outliers. Although they are shown in red lines, the graph is not very clear. The outliers should be shown in the line graph.	Thank you for highlighting the issue of identifying outliers in the datasets, with respect to the 95% confidence bounds. Whilst we believe that Figure 5 is very effective in identifying the presence of outliers, we do recognise the importance of showing the outliers on the raw	Page 24, Line no. 397-398 Figure 6 updated

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		settlement-time series graph. Therefore, in response to the reviewer's comment, we have inserted four additional figures into Figure 6 (a, b, e, f), which show the settlement-time series before PCA with example outliers indicated within the circles on the graphs (Figure 6a and 6b). Figure 6e and 6f show zoomed in indication of outlier removal for these specific areas as an example.	
7	According to Table 4 and Table 5, the filtering and smoothing effects of different methods are not significantly different. What do the authors think are the main differences among different filtering methods?	The reviewer correctly highlights that there is a high degree of similarity between the performance of the filter methods used. Having explored numerous filtering approaches, the four filters which we have presented results for were the most consistent. However, based on the RMSE and NMSE values, none of these four filtering approaches can be considered as being consistently most effective.	Page 27, Line no. 468-474 "*It is clear that none of the four filtering approaches used outperformed the others. These approaches were selected due to their high levels of consistency in terms of NMSE and RMSE values. Subtle variations are apparent (e.g. PM-E01 in Table 4), whereby Savitsky-Golay appeared to be slightly better. Whereas for PM-E02 (Table 5), the*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		To address this, the authors have inserted some new text into the manuscript.	*Zero-Phase filtering method appeared to perform better. This highlights that without implementing the four filtering approaches on the settlement data, it would not be possible to fully assess which approach is more preferable.*"
8	The amount of data in this paper is too small to reflect the filtering effect. Also, why not use machine learning modeling to justify data preprocessing? Thus, this paper is incomplete and elementary.	The authors appreciate the reviewer's comment. To clarify, this study focusses on addressing the practicality of data acquisition and data-driven modelling in modern -day geotechnical engineering. Previous studies involving the use of data analysis on geotechnical settlement data were carried out on purpose-built test embankments with a high density of instrumentation. For the majority of commercial civil engineering projects involving embankment construction, it is very rare for large and high-quality settlement data sets to be collected from a high number of instruments due to cost	N/A

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		restrictions. This study challenged the applicability of data analytical methods typically used on big data sets, but on a comparatively small settlement dataset which contained various errors (due to the manual method of data collection) and poor levels of quantisation. The resulting analysis provides confidence to geotechnical engineers for making informed judgements regarding ground response to embankment construction on site. Given these data characteristics, the authors believe that the analysis presented in the paper is far from being elemental. With regards to the reviewer's comment on machine learning, this would require data sets that are much larger than what was available for this study. In addition, the purpose of performing machine learning modelling is to identify trends within	

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		data sets, which must be free of as many outliers as possible. Additionally, machine learning does not consider underlying engineering knowledge regarding the dynamics of the embankment settlement, which is essential for successful outlier detection and removal. Based on the number of outliers present within the raw time-based settlement data, the application of machine learning algorithms was not applicable for this study. We trust that this response satisfies the reviewer.	
Reviewer 2			
1	Very interesting paper with very relevant subject matter. The main issue I have with the paper is you do not address the impact (benefit/consequence) this data correction procedure has on your calculated consolidation nor do you address the risk of falsely removing	The authors would like to thank the reviewer for such a kind comment on our work – we are glad to hear that it is of interest. To re-emphasise the impact of this study, the data analysis performed is	Page 2, Line 19-21 (Abstract) "*The findings from this study would be useful to site engineers regarding complex decision-making relating to ground response due to embankment construction.*"

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
	correct data. I think a discussion on both would add greatly to the paper.	beneficial in that it can provide site engineers with improved short-term information about embankment settlement dynamics. Without this, data outliers can obscure the true data trends. We fully acknowledge the reviewer's comment regarding incorrect removal of data points – this can be a challenging issue and is sometimes inevitable. However, the risk of accidentally removing correct data points can be minimised by using experienced engineering judgement. We discussed this in Line 487-490**,** and we have added further clarification. For this particular study, geotechnical knowledge of how the soils present beneath the embankment were formed and their mechanical behaviour meant that it was not possible for them to experience	Page 32, Line no. 506-516 *"However, the removal of settlement outliers and data cleaning by filtering and smoothing methods is often not a straightforward process. The challenging issue with data cleaning is, in general, the risk of accidentally removing correct data points. Although this is sometimes inevitable, it can be minimised by maintaining human-in-the-loop to adjust filter parameters for an optimised result. Whilst the RMSE and NMSE metrics that have been used are useful for informing the filter performance, these are based on experienced engineering judgement. For this particular study, geotechnical knowledge of how the soils beneath the embankment were formed and their mechanical behaviour meant that it was*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		heave during embankment loading. However, cleaning data by removing all data points that suggest heave is an act of over conservatism, whereby there would have been a risk of falsely removing correct data. To reflect the above responses, the authors have inserted some additional text into the manuscript.	*not possible for them to experience heave during embankment loading. However, cleaning data by removing all data points that suggest heave is an act of over-conservatism, whereby there would have been a risk of falsely removing correct data."* Page 36, Line no. 592-595 *"Without this, data outliers can obscure the true data trends. Therefore, the data analysis performed in this paper was also beneficial in providing site engineers with improved short-term information concerning embankment settlement dynamics."*
2	I would also like to see greater discussion on Figures 7 and 8, as we can see clear differences between the original data and filtered data but it is not clear why we should trust this data more.	Thank you very much for raising this comment. As the reviewer has identified, there is a slight variance between the raw data and the filtered-smoothed data curves. Based	Page 27, Line no. 459-466 *"Based on soil consolidation theory, it is expected that for any given vertical load applied to the soil, its settlement behaviour will be characterised by a*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		on basic soil consolidation theory, it is expected that for any given vertical load applied to the soil, its settlement behaviour will be characterised by a smooth exponential curve. Also, there will be a maximum rate at which settlement will occur (based on soil material properties such as stiffness and permeability). Some of the raw data points will indicate that settlement occurred at a higher rate than this, due to the presence of data noise/outliers. The resolution of the raw settlement data was 1 mm. However, filtering results showed values with a higher level of resolution. Therefore, filtering results showed a smoother trend compared with the raw data and therefore more characteristic of field behaviour. The authors have inserted some new text within the manuscript to reflect the	*smooth exponential curve [50]. Moreover, based on soil material properties such as stiffness and permeability, there will be a maximum rate at which settlement will occur. Therefore, the variance in the raw data and filtered-smoothed data curves is due to 1) presence of noise in the data after removal of outliers and 2) resolution of the raw settlement data was 1 mm whereas filtering results showed values with a higher level of resolution. Therefore, filtering results showed a smoother trend compared with the raw data and therefore more characteristic of field behaviour."*

Comment no.	Editor / Reviewer comment	Author response	Amendments to submission
		responses provided above. We trust this now satisfies the reviewer.	

A case study using PCA and signal processing techniques for processing time-based construction settlement data of road embankments

Faisal Siddiqui [1*], Paul Sargent [1], Gary Montague [2]

[1] School of Computing, Engineering and Digital Technologies, Middlesbrough, Tees Valley, TS1 3BX, UK

[2] School of Health and Life Sciences, Teesside University, Middlesbrough, Tees Valley, TS1 3BX, UK

[*] Corresponding author

Faisal Siddiqui (E: f.siddiqui@tees.ac.uk); Paul Sargent (E: p.sargent@tees.ac.uk); Gary Montague (E: g.montague@tees.ac.uk)

Highlights

- PCA was successfully applied in removing outliers from dynamic settlement data.
- Comparison of data smoothing techniques for cleaning noisy settlement data.
- Settlement and pore water pressure data were compared to verify dynamic behaviour.
- Information content from commercial geotechnical datasets were maximised.

A case study using PCA and signal processing techniques for processing time-based construction settlement data of road embankments

Faisal Siddiqui [1*], Paul Sargent [1], Gary Montague [2]

[1] School of Computing, Engineering and Digital Technologies, Middlesbrough, Tees Valley, TS1 3BX, UK

[2] School of Health and Life Sciences, Teesside University, Middlesbrough, Tees Valley, TS1 3BX, UK

[*] Corresponding author

Faisal Siddiqui (E: f.siddiqui@tees.ac.uk)

Paul Sargent (E: p.sargent@tees.ac.uk)

Gary Montague (E: g.montague@tees.ac.uk)

Abstract

Instrumentation is beneficial in civil engineering for monitoring structures during their construction and operation. The data collected can be used to observe real-time response and develop data-driven models for predicting future behaviour. However, a limited number of sensors are usually used for on-site civil engineering construction due to cost restrictions and practicalities. This results in relatively small raw datasets, which often contain errors and anomalies. Interpreting and making judicious use of the available dataset for developing reliable predictive model represents a significant challenge. Therefore, it is essential to pre-process and clean the data for improving their quality. To date, little investigation has been performed in the application of such data cleaning methods to geotechnical engineering datasets collected from full-scale sites. The purpose of this study is to apply simple and effective data pre-processing techniques to site-data collected from a highway embankment constructed on a sequence of soil layers of different physical make-up and non-linear consolidation characteristics. Various cleaning methods were applied to magnetic extensometer data collected for monitoring settlement within foundation soils beneath the embankment. PCA was used to explore raw data, identify and remove outliers. Numerous filtering and smoothing methods were used to clean noise in the data and their results were further compared using RMSE and NMSE. The methods adopted for data pre-processing and cleaning proved very effective for capturing the raw settlement behaviour on site. The findings from this study would be useful to site engineers regarding complex decision-making relating to ground response due to embankment construction. This also has positive prospects for developing dynamic prediction models for embankment settlement.

Keywords: Embankment Construction; Soil Settlement; Data Pre-processing; Principal Component Analysis; Signal Processing; Data Cleaning

Abbreviations: PCA – Principal Component Analysis; RMSE – Root Mean Square Error; NMSE – Normalised Mean Square Error; ME – Magnetic Extensometer; VWP – Vibrating Wire Piezometer

1. Introduction

Instrumentation networks are generally installed within or immediately adjacent to geotechnical structures such as retaining walls, embankments and structural foundations, in order to closely monitor their stability during and after construction. This ensures their full compliance with serviceability limit state design. Another benefit of the monitoring datasets is that they enable quick identification of any unexpected structural behaviour, thereby providing an early warning to facilitate timely remedial actions [1]. One of the largest challenges for civil engineers is to move from the data-rich environment, resulting from a diversity of measurements, to one that is information-rich [2]. This is compounded by practical difficulties such as site access, which may be inhibited due to unfavourable weather and/or ground conditions. Additionally, most construction projects have financial constraints associated with the number of instruments that can be installed and the method of data collection (i.e. digital data logging vs manual readings). These factors result in generating smaller monitoring datasets compared with those for higher profile construction projects and more academic research-based studies, which have significantly more funds for installing advanced and sophisticated instrumentation networks.

The capacity of the UK's civil engineering infrastructure is growing to accommodate increasing numbers of road and railway users. This includes the construction of new or expansion of existing highways and railways [3,4]. Earth embankment structures are widely used to enable construction of these pieces of infrastructure. Magnetic extensometers (ME) and vibrating wire piezometers (VWP) are traditionally installed within the embankments to monitor the non-linear settlement and pore water pressure response of the foundation soils during construction [5]. These measurements are used to compare predicted versus observed calculations for consolidation of the embankment and underlying strata. Field and laboratory

53 based predictions are generally performed by using numerical methods such as one-
54 dimensional consolidation, the Terzaghi method and finite element analysis (FEA) [6,7].

55 ME and VWP instruments have recently been used by Zheng et al. [8] to study the
56 settlement behaviour of an embankment constructed in Ballina (NSW, Australia) on a soft
57 lightly over-consolidated clay of high to extremely high plasticity. The settlement observed
58 during and after construction was higher than FEA predictions based on field and laboratory
59 tests, which would lead to higher costs associated with "*surcharge, stripping and spoiling*
60 *during construction*" [8]. This demonstrates that whilst FEA is a powerful tool in design, it
61 does not consistently provide accurate predictions of site behaviour. For validation, they
62 require accurate site information and soil parameters (e.g. small strain stiffness from triaxial
63 testing) that can be derived from expensive laboratory testing techniques [5]. An alternative
64 method is to calibrate FEA models by updating soil parameters using statistical methods to
65 improve prediction capability. This requires on-site measurement data through the installation
66 of ME and VWP instruments [8–10].

67 Data analysis, modelling and forecasting can be utilised to study the long-term
68 behaviour of a highway embankment by predicting settlement from historical monitoring data.
69 A classic data-driven process includes (i) operation evaluation; (ii) data acquisition and data
70 pre-processing; (iii) feature selection and (iv) data-driven model development using statistical
71 and machine learning methods [11]. In operation evaluation, the mechanical behaviour
72 mechanisms and material strength parameters that would affect the behaviour of the structure
73 are determined. This knowledge is then used to select the critical locations for installing
74 monitoring instrumentation in the structure to effectively capture changes in material and
75 structural behaviours. It is also essential to determine the most appropriate frequency at which
76 data points are acquired, along with the robustness and accuracy of the instruments to be used
77 [12]. Data generated from the instruments can be collected either manually by site engineers

or digitally via data loggers, whereby data can either be downloaded manually or remotely via Bluetooth or 4G internet connection [13].

Data pre-processing is a preliminary step in developing data-driven models, which improves the data quality [14]. Data analytics such as machine learning and statistical algorithms used for developing data-driven models are sensitive to data quality. Thus, the presence of noise and outliers in the datasets may lead to incorrect interpretations [15]. The improvement in raw data quality brings completeness, accuracy and confidence in the dataset and reliability in data analysis results [16,17]. Therefore, correctly removing outliers and reducing noise in the datasets is of paramount importance. Pre-processing also involves other elements of data cleaning, specifically the process of identifying and repairing duplicates, missing values in the dataset [18].

There are a limited number of studies where soil parameters have been defined based on data collected from instruments or sensors, compared with those determined from laboratory tests [19]. While various case studies have involved data analysis for predicting soil properties and behaviour [20,21], the strategy of data cleaning in terms of geotechnical instrumentation is currently not well defined. The quality of data collected from active monitoring sites can be adversely affected by factors including sensor calibration errors, limited sensor resolution, effects from the environment (e.g. heat, moisture, atmospheric pressure) on sensors and human error associated with manual measurements. Therefore, it is likely that datasets can get corrupted with noise and contain anomalies.

The aim of this study is to utilise data gathered from a recently completed highway construction project to demonstrate the use of data analytical methods for improving data quality and its interpretation for enhanced prediction of embankment settlement. The instruments used for data collection were in accordance with UK engineering standards [13], and installed to monitor settlement and pore water pressures in the soil strata underlying an

103 embankment during the construction phase. These facilitated construction of the finalised
104 embankment drainage scheme and road surfacing, whereby once excess pore water pressure
105 dissipated, and ground movement had stabilised to reduce final differential settlement of the
106 embankment. Since the data was collected only during the construction phase, the objective
107 was to make judicious use of the short-term dataset by enhancing the quality through the
108 application of cleaning methods. The paper presents a methodology for pre-processing in
109 preparing geotechnical monitoring instrumentation data for a highway embankment case study,
110 with a view to developing a data-driven model and investigate statistical methods to enhance
111 quality and make better settlement predictions. Commonly used PCA, filtering and smoothing
112 methods were adopted for processing the data to encourage their use as an industrial standard
113 for processing raw geotechnical data. These procedures could be more generally applicable for
114 data cleaning of other geotechnical instrumentation data, which would improve the efficiency.
115 Although the concepts of data pre-processing and cleaning used are well established in other
116 domains, they are still in their infancy in the field of geotechnical engineering.

117 The structure of the paper is organised as follows. Section 2 presents the case study
118 used for this paper, which includes a description of the instrumentation used, their location and
119 data acquisition frequency. Section 3 provides a theoretical background on data pre-processing
120 in preparation for data analysis, specifically focusing on PCA, signal processing-based filtering
121 and smoothing that have been utilised in this paper for exploring and processing raw data.
122 Section 4 presents the methodology used for investigating and improving the quality of raw
123 embankment settlement data, including: (i) PCA for initial investigation and detecting outliers;
124 (ii) filtering and smoothing to remove noise from the data and improve their quality. Section 5
125 discusses the data pre-processing results and their statistical comparison using RMSE and
126 NMSE. Section 6 discusses the overall result and their credibility. Section 7 provides a

summary of key findings from the paper and recommendations for future processing of raw geotechnical data.

2. Case study – Morpeth Northern Bypass, Northumberland, UK

Morpeth Northern Bypass (A197) is a 3.8km highway, located immediately north of Morpeth town centre in Northumberland, UK. Construction of the bypass started in 2015 and has been in operation since 4th April 2017. Numerous geotechnical structures were constructed as part of the bypass, including four new earth embankments and multiple bridge abutment foundations supported by reinforced soil blocks. Each structure was appropriately instrumented to monitor their on-site behaviour during construction. For this study the new earth embankment constructed at Pegswood Moor was considered, which is located towards the eastern limit of the bypass (see Figure 1). This embankment was selected as the focus for this study as 450-500mm of settlement was predicted based on the raw ground investigation data, which was the largest out of the settlement predictions made for all four embankments on the scheme.

The superficial geological conditions which underlie the bypass are largely glacial deposits; comprising a stiff upper over-consolidated till, a middle unit of sands and gravels, which is further underlain by a lower till. Findings from ground investigation surveys undertaken on the site prior to construction of the bypass revealed that the base of the upper till is soft, of higher plasticity and commonly laminated. The underlying solid geology is chiefly characterised by Carboniferous strata, namely the Coal Measures formation which comprises interbedded sandstones, siltstones, mudstones, seatearth and coal.

The Pegswood Moor embankment was constructed from cohesive engineering fill, with a maximum height of 7m and maximum width of 52m. The side slopes of the embankment had a maximum gradient of 1V in 2.5H. Pre-fabricated vertical drains were installed within the

151 foundation soils to enable faster dissipation of pore water and hence consolidation during the
152 embankment construction. Four instruments were installed within the foundation soils situated
153 beneath the embankment to monitor settlement and pore water pressure responses to loading:
154 two ME's (PM-E01 and PM-P02) and two VWP's (PM-P01 and PM-P02). Figure 2 presents a
155 cross section of the embankment geometry, instrumentation installed and geological
156 information obtained from locally drilled exploratory boreholes. The specific depths of
157 instrumentation sensors and soil descriptions at these corresponding depths are detailed in
158 Tables 1 and 2.

159 The methods followed for installing the instrumentation was based on guidance
160 provided by the ICE [13]. Brief descriptions of the ME's and VWP's are provided below.

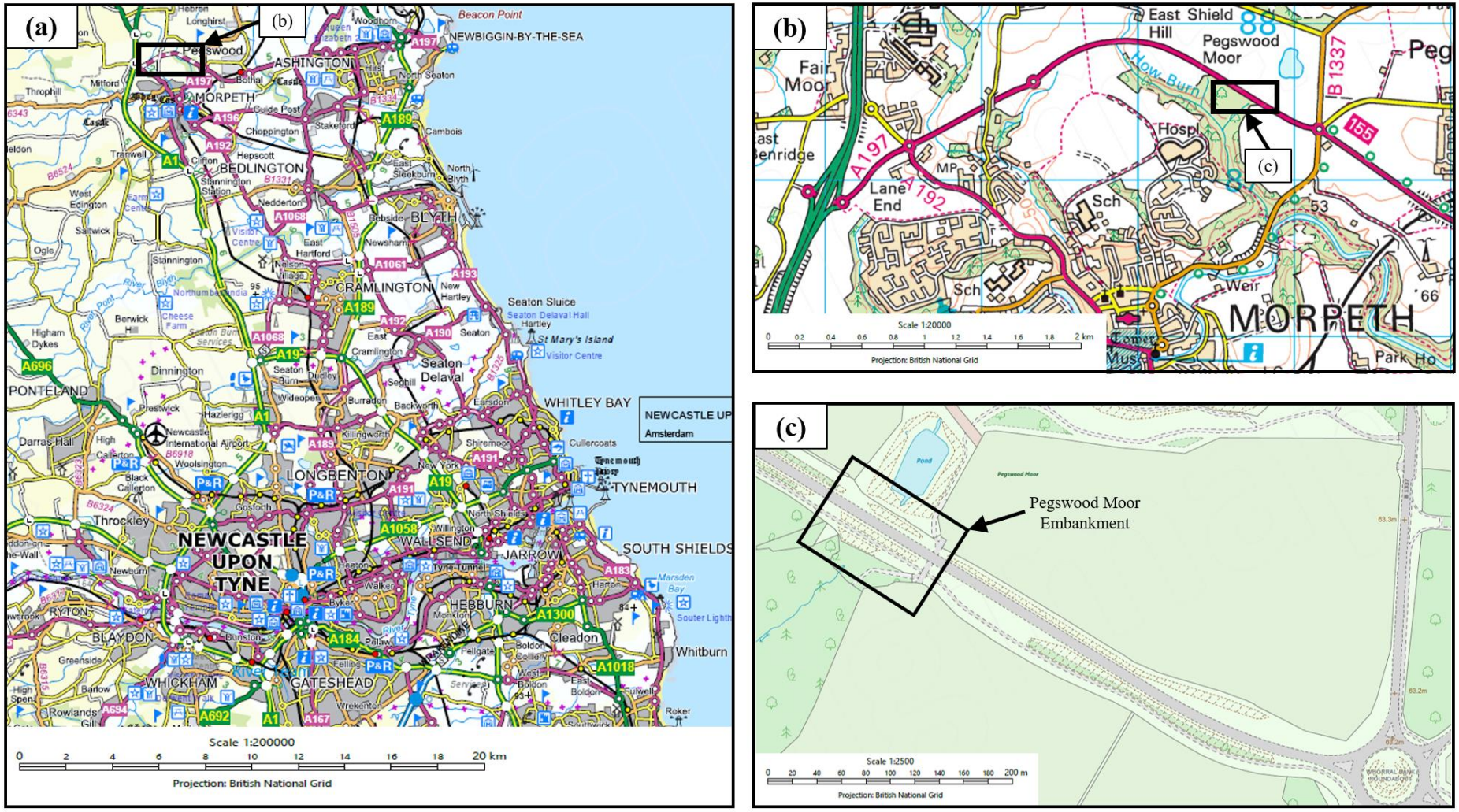


Figure 1: Site location (courtesy of Edina Digimaps [22]).

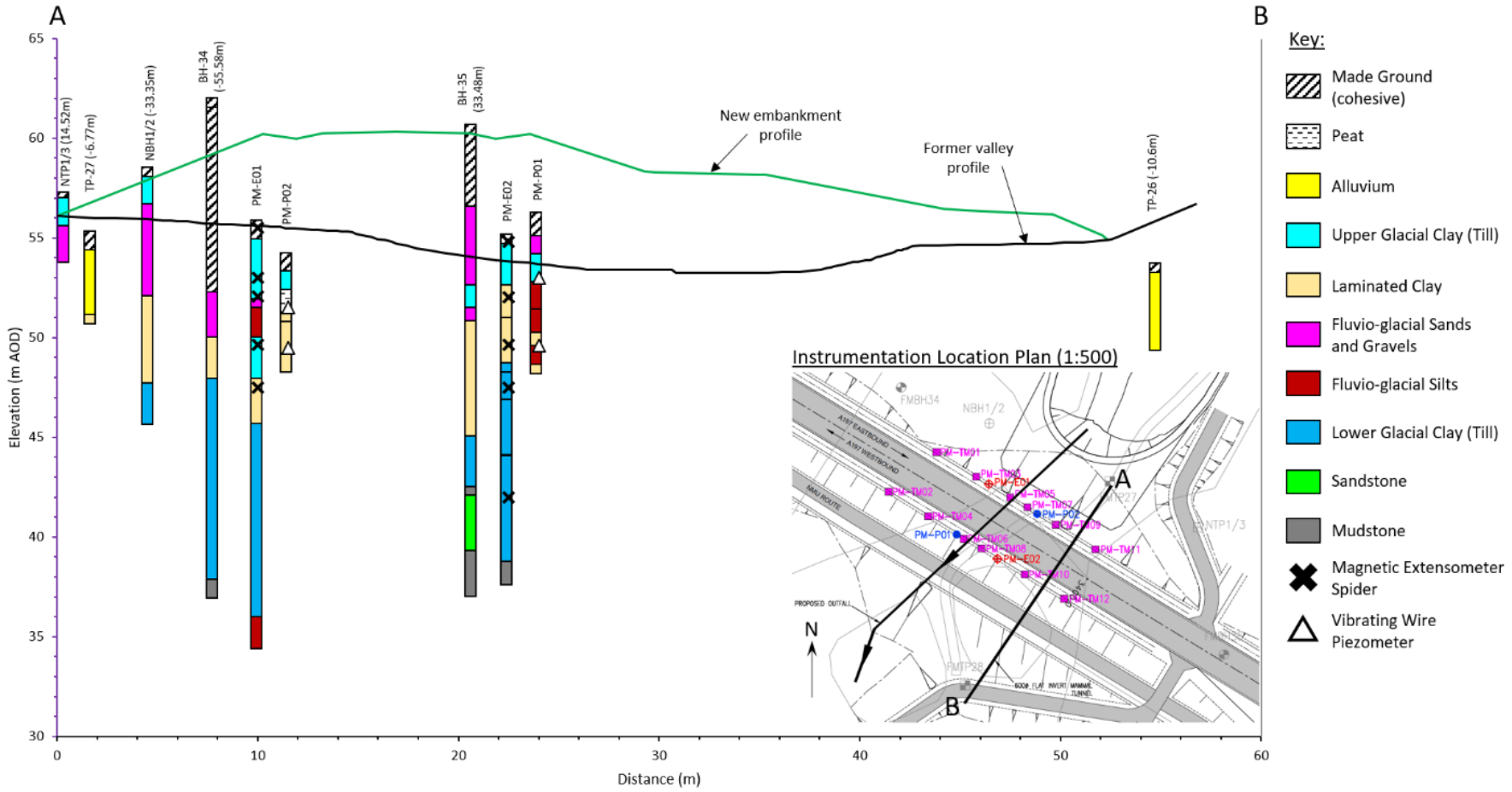


Figure 2: Pegswood Moor embankment geometry, with instrumentation installation and geology (borehole) information superimposed.

Table 1: Instrumentation details and soil information recorded for magnetic extensometer locations.

Instrument reference	Instrument sensor reference	Installation depth (m bgl)	Soil description	Soil depth (m bgl)
PM-E01	Plate Magnet 1	0	MADE GROUND (cohesive)	0 – 1
	Spider 4	2	Laminated CLAY	1 – 3.8
	Spider 3	4.6	Sandy SILT	4.5 – 6.10
	Spider 2	6.1	Sandy laminated CLAY	6.1 – 7
	Spider 1	7.9	Soft to firm silty laminated CLAY	7.8 – 10.8
	Base Magnet	20	Weak MUDSTONE	19.9 – 21.45
PM-E02	Plate Magnet 1	0	MADE GROUND (granular)	0 – 0.3
	Spider 4	2.7	Soft sandy gravelly CLAY	2.5 – 4
	Spider 3	4.2	Soft slightly sandy silty CLAY	4 – 6.6
	Spider 2	7.2	Firm to stiff sandy silty CLAY	7 – 8.4
	Spider 1	8.6	Firm to stiff sandy silty CLAY	8.4 – 11
	Base Magnet	15.5	Weak MUDSTONE	16.5 – 17.5

Table 2: Instrumentation details and soil information recorded for vibrating wire piezometer locations.

Instrument reference	Instrument sensor reference	Installation depth (m bgl)	Soil description	Soil depth (m bgl)
PM-P01	P01-B	3	Soft silty sandy laminated CLAY	2 – 3.5
	P01-A	6.4	Soft slightly silty laminated CLAY	6.1 – 6.6
PM-P02	P02-B	2.5	Soft firm silty clayey PEAT	1.8 – 3
	P02-A	4.5	Soft silty laminated CLAY	3.5 – 5.1

2.1.Instrumentation

2.1.1. Magnetic Extensometer (ME)

These instruments are widely used to measure embankment settlement, whereby vertical movements are measured with respect to a magnet placed at the base of the borehole within a stiff stratum, which is unlikely to be impacted by embankment construction. The instrument comprises a number of settlement targets called 'spiders' and plate magnets, which are installed within the walls of the borehole in compressible soil layers of interest. Readings are taken manually by using a magnetic sensor that is attached to a measuring tape, which has an audio and visual indicator. The probe is lowered down the borehole and when it reaches depths which coincide with the positions of the spiders and plate magnets, the indicators are activated [23]. The 'soil instruments' manual states that settlement can be calculated using the probe data, using Equation 1:

$$Settlement = (D_0 - M_0) - (D_i - M_i); \quad i = 1, 2, \ldots n \qquad (1)$$

where D_0 is the initial depth reading from the reference point (ordinance datum i.e. mean sea level [mAOD]) to the base magnet; M_0 is the initial reading from the reference point (ordinance datum i.e. mean sea level [mAOD]) to the spider or plate magnet; D_i is the subsequent reading from the reference point to the base magnet; M_i is the subsequent reading from the reference point to the spider or plate magnet and n is the number of data points [24].

2.1.2. Vibrating wire piezometer (VWP)

VWP's are widely used to monitor groundwater pressures during and post construction. They are stable, easy to use and do not interfere with construction activities. VWP's consist of a vibrating wire pressure transducer which is connected to a data logger. When pore water pressure acts on the diaphragm attached to the transducer by a sensing wire, it deflects and induces wire tension. As the level of tension changes in response to changing pore water pressures, this results in changes to the frequency of vibration [25]. Pore water pressure can thereby be calculated by using Equation 2 as follows:

$$P = K(f^2 - f_0^2) \qquad (2)$$

where P is the pore water pressure; K is a constant provided by the manufacturer; f *and* f_0 are current and base frequency respectively [23].

2.2.Embankment Construction

The embankment was chiefly constructed from Class 2 general cohesive fill, derived from the Upper Glacial Till recovered from cuttings formed as part of scheme bulk earthworks. A 600 mm thick layer of selected uniformly graded granular Class 6C material was placed at the base of the embankment to facilitate construction and provide active drainage. Based on one-dimensional consolidation (oedometer) tests that were performed on clay samples obtained from the original ground investigation, the rate of embankment construction was estimated to

be 1m of fill every 2 weeks; whereby 2 days would be required to place the fill followed by a holding period of 12 days. This was considered sufficient to allow dissipation of excess pore water pressures generated due to fill placement, in addition to foundation soils having the ability to regain effective strength and achieve an adequate state of consolidation before applying additional embankment fill. To reduce the overall construction period, a network of pre-fabricated vertical band drains (each drain being 100mm wide, 4mm thick, up to 8m long and at 0.75m spacings) was installed in the foundation soils to enable faster rates of excess pore water pressure dissipation [26]. By coupling the band drains with the Class 6C layer at the embankment base, this resulted in reducing the final hold period post-construction from 45 weeks to 8 weeks and a limiting residual settlement of less than 25mm.

The VWP and ME monitoring datasets also had the purpose of confirming the sufficiency of the reduced hold periods for consolidation of the foundation soil layers. The settlement data was further used to determine the final settlement of the soil layers using Asaoka method, to enable comparisons with estimates made from oedometer tests [27]. Although the Asaoka method is a widely used approach for predicting final settlement, it cannot be used to study the ground settlement behaviour during early phases of construction [26].

3. Theoretical Background

Raw data requires initial exploration and data quality assessment to check their reliability for data-driven modelling [28]. During civil engineering structural monitoring, Li et al. [29] highlighted how missing and abnormal data can have negative impacts on dataset quality, suggesting that data cleaning is required to mitigate against this. According to Cernuda and Krishnan et al. [14,18], data cleaning generally incorporates the following data enhancement approaches: (i) interpolation to deduce missing values, (ii) data deduplication to reduce data volume and (iii) outlier detection. There is often a requirement for “user

supervision" in data cleaning, which ranges from the definition of data quality rules to manual identification and fixation of errors [30]. Therefore, the nature of data cleaning can be iterative and involve user expertise, resulting in a human-in-the-loop cleaning system [18].

In terms of time-series data, Wang and Wang [31] suggested a two-stepped process for data cleaning – the first step is to detect and eliminate outliers in the dataset. The second step involves data cleaning using any of the following algorithms: (i) smoothing-based, (ii) constraint-based or (iii) statistics-based.

3.1. Principal Component Analysis (PCA)

PCA is an unsupervised exploratory data analysis technique, whereby multiple correlated variables in datasets are combined linearly to reduce them to uncorrelated variables called 'principal components'. These principal components are a set of eigenvectors, which are sorted in descending order of their eigenvalues [32]. Eigenvalues for each principal component represent the amount of variance between the variables, which are important to determine the significant number of principal components by considering those with eigenvalues greater than average eigenvalues. The data points corresponding to these principal components are termed 'scores', which can be plotted against each other to show the position of the data sample in the principal component space [33]. For data exploration, characteristics of data can be qualitatively investigated by using score plots, which show the position of the data sample in principal component space [34]. PCA allows exploration and investigation of multivariate datasets in a reduced dimension, while preserving the characteristics of the data in principal components [35]. Further detailed description of PCA is presented by Jolliffe and Cadima [36].

In PCA, data scaling is applied to ensure that the influence of all variables is equal in determining the number of principal components. Auto-scaling is a commonly used method, whereby data points are centred and scaled by subtracting the variable mean value and dividing

by the variable standard deviation. After scaling, a significant number of principal components will have an eigenvalue greater than 1 [37].

PCA provides Hotelling T^2-statistic and Q-statistic results, which can be plotted against each other to produce an influence plot. This is subsequently used for identifying anomalies in the dataset by identifying data points which are outside the confidence bounds of Hotelling T^2-statistic and Q-statistic [34,37].

3.2. *Signal processing*

In civil engineering, signal processing approaches have previously been used for system identification by utilising many time-series datasets obtained from structures [38]. Signal processing-based filtering and smoothing can be utilised to pre-process and improve the signal quality by cleaning noisy data [31,39]. For time-series datasets, reasons for poor data quality can include environmental effects (e.g. temperature, barometric pressure), inaccurate calibration of sensors and subsequent drift [40]. Environmental effects could result in high frequency noise within datasets although this may be compounded by more severe deviations in the signal, such as electrical spikes and human errors associated with manual data entry. This latter type of signal corruption may be referred to as an 'outlier' [14].

For data filtering and smoothing, noise is generally removed by using moving-average or low-pass filtering algorithms. However, there are many data smoothing techniques available for use [31,41]. Generally, the moving average method is the simplest and most widely used [31]. If there are outliers in the time-series data that might distort the smoothing process, the weighted moving average method is preferable [42]. Savitzky-Golay is a finite impulse response (FIR) smoothing filter based on the least-square method, which involves fitting a polynomial curve to a given range of data points [43]. Zero-phase 'time reversal' filtering can also be applied, which involves filtering time series data twice by flipping the dataset and can

be used in pre-processing of archival data [40]. In this method, different filtering methods can be utilised; whereby the most basic and common method is Butterworth [11].

Processing and cleaning datasets by filtering and smoothing requires filter parameter selection to achieve the desired result with minimal discrepancy between the filtered and original data [44]. Unfortunately, such filter parameters cannot be predetermined before implementation. Hence, an iterative procedure is required to select the most appropriate values and achieve the desired results. In comparing the results of different filter parameters, it can be useful to have quantitative measures of their effectiveness. Statistical indicators include Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) [29], [40].

4. Data pre-processing methodology

The data pre-processing methodology comprised numerous steps, whereby these and the resulting data conditions are presented in Table 3.

Table 3: Summary of data pre-processing methodology.

Stage no.	Process	Outcome
1	Initial Data Investigation	Duplicate data points removed; Raw Data with outliers and noise
2	Unsupervised Data Exploration (Principal Component Analysis)	Qualitative understanding of data; Outlier detection and removal
3	Data Quality Enhancement (Moving-average; Gaussian-weighted; Savitzky-Golay; Zero-phase)	Realise data trend; Filtered and smoothed data
4	Filter Comparison and Validation	Data credibility

After removal of duplicate data points from the ME dataset, PCA was performed for unsupervised exploration of the data using a graphical user interface-based PCA toolbox in MATLAB ver. 2019b [34,45]. Score values of the first two principal components captured >90% of the overall data variation and were therefore sufficient to investigate the settlement and confirm data reliability. Furthermore, the PCA-based metrics of Q statistic and Hotelling-

T^2 were used to identify outliers and remove them from the dataset. For simplicity, data that was removed was replaced by using linear interpolation based on the data points either side. To further confirm whether ME data points were indeed outliers, comparison was made with measurements of pore water pressure from adjacent VWP's; whereby rises in pore water pressures are generally correlated with increases in settlement due to embankment construction.

Following outlier removal, filtering and smoothing methods were then used to remove high frequency noise and identify significant trends in the data. Filtering and smoothing were performed by using MATLAB ver. 2019b [44,45]. The following methods were chosen based on an overall balance between their simplicity and effectiveness: (1) moving average smoothing, (2) Savitzky-Golay filtering, (3) Zero phase filtering and (4) Gaussian-weighted moving average method. Filter parameters used in these methods were tuned iteratively based on engineering judgement to best capture the system response in terms of smooth settlement curves. The results were compared by visual inspection and applying quantitative statistical analysis: specifically (1) Root Mean Square Error (RMSE) and (2) Normalised Mean Square Error (NMSE). To confirm the effectiveness of the filtering, PCA was reapplied to the filtered data to confirm that no outliers were present.

5. Results

5.1. Unsupervised exploration of data and outlier detection

ME data had 5 variables representing different spiders and plate magnets installed within soil layers beneath the embankment, to which PCA was applied. Firstly, all variables were normalised by autoscaling where data points were mean centred and further divided by their standard deviation [35,37].

The number of principal components required to describe major settlement variation of all soil layers was determined to remove noise. The dataset was explored in PCA by plotting the eigenvalue against the principal component to determine the number of components having a value >1. Kaiser's rule (AEC) states that components with Eigenvalues >1 are relevant, whereas those <1 describe noise. A more conservative threshold value would be 0.7, as specified by CAEC [34]. Figures 3a-3b show Eigenvalue vs Principal Component plots for PM-E01 and PM-E02, whereby the outputs from these graphs can determine whether further principal components should be considered. The blue and red horizontal lines correspond to the AEC and CAEC limits respectively. It can be seen that one principal component is sufficient to capture significant variation of all soil layers and accounts for >90% of the observed variance. The first two components account for 98.78% variance in PM-E01 and 99.63% in PM-E02 explaining >95% of the total data characteristics information [35]. When attempting to observe overall system behaviour, subtle variations may become important and therefore PCA selection should not be limited to a 90% threshold. Hence, a set of RMSECV cross-validation plots were also considered for PM-E01 and PM-E02 datasets in Figures 3c-3d [37].

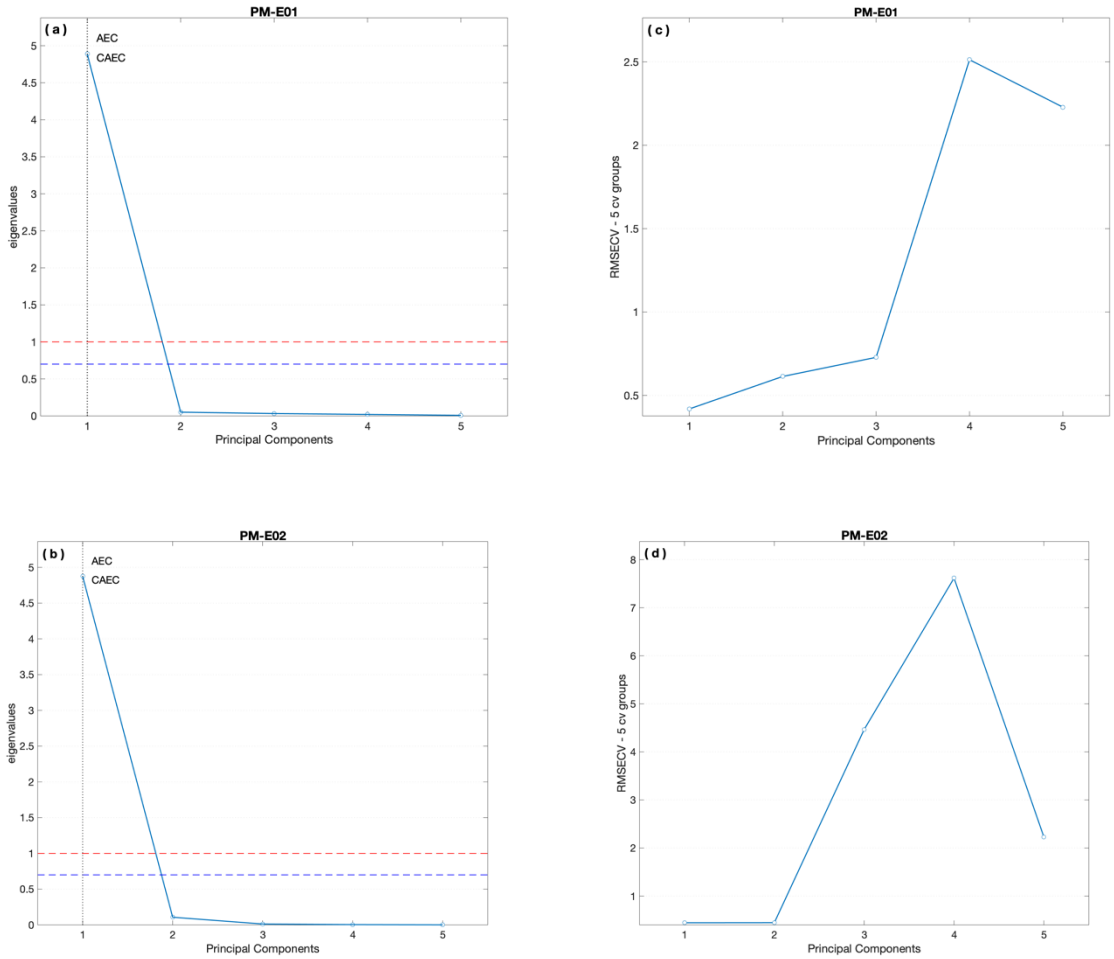


Figure 3: Eigenvalue (left) and RMSECV (right) for (a and c) PM-E01 and (b and d) PM-E02.

It is observed in Figure 3c that PM-E01 shows a gradual increase in the first three principal components and a sudden increase in the fourth component. In Figure 3d for PM-E02, the first two principal components remain constant with a sudden increase for the third principal component. These plots suggest that it is necessary to consider two to three principal components to explore the settlement behaviour of the system. The other principal components are considered to be describing noise within the system and therefore can be discounted.

The threshold set by AEC suggests that one principal component is sufficient for capturing the underlying behaviour of the system. Whereas, the RMSECV cross-validation approach suggests that considering additional principal components would be beneficial. Hence, this study considers principal components 1 and 2 in the subsequent analysis to observe

clustering of settlement data for each 1m fill increment of the embankment, as shown in Figure 4.

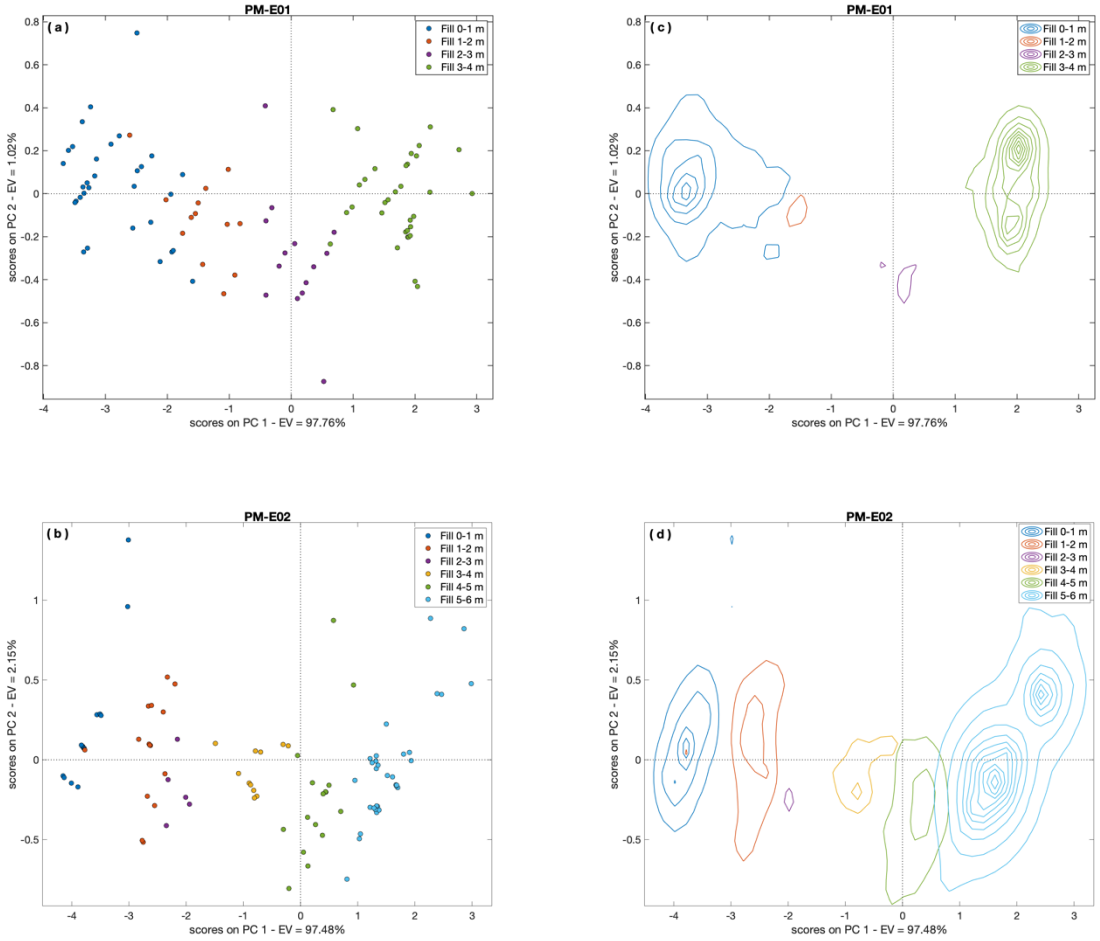
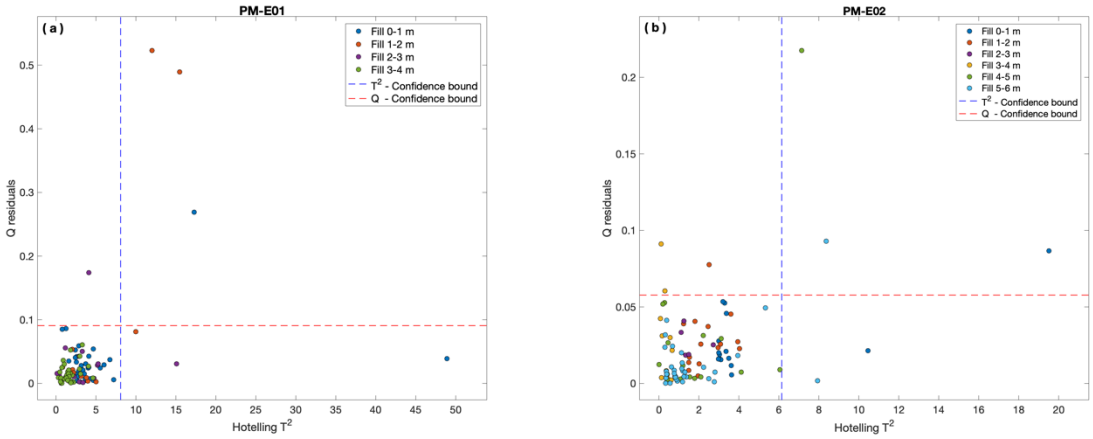


Figure 4: Score Plot - Scatter (left) and Contour (right) for (a and c) PM-E01 and (b and d) PM-E02.

Figures 4a and 4c demonstrate that the first and second principal components show a clear separation between different 1m increments in embankment fill level. These figures also show distinct data clusters for each fill increment. These clusters provide confidence that the raw data, with suitable pre-processing, are sufficiently rich in information to distinguish the settlement response between different embankment fill increments. Therefore, this data can be further processed and prepared for subsequent modelling such as shallow neural networks. However, for PM-E02 data presented in Figure 4d, a slight overlap can be observed between

365 the clusters relating to fill levels 59 and 60 mAOD. This can be expected due to the incremental
366 settlement reducing as the embankment is constructed, as a result of progressive foundation
367 soil consolidation. Thus, in the latter stages of embankment construction, more overlapping of
368 data clusters would be considered likely.

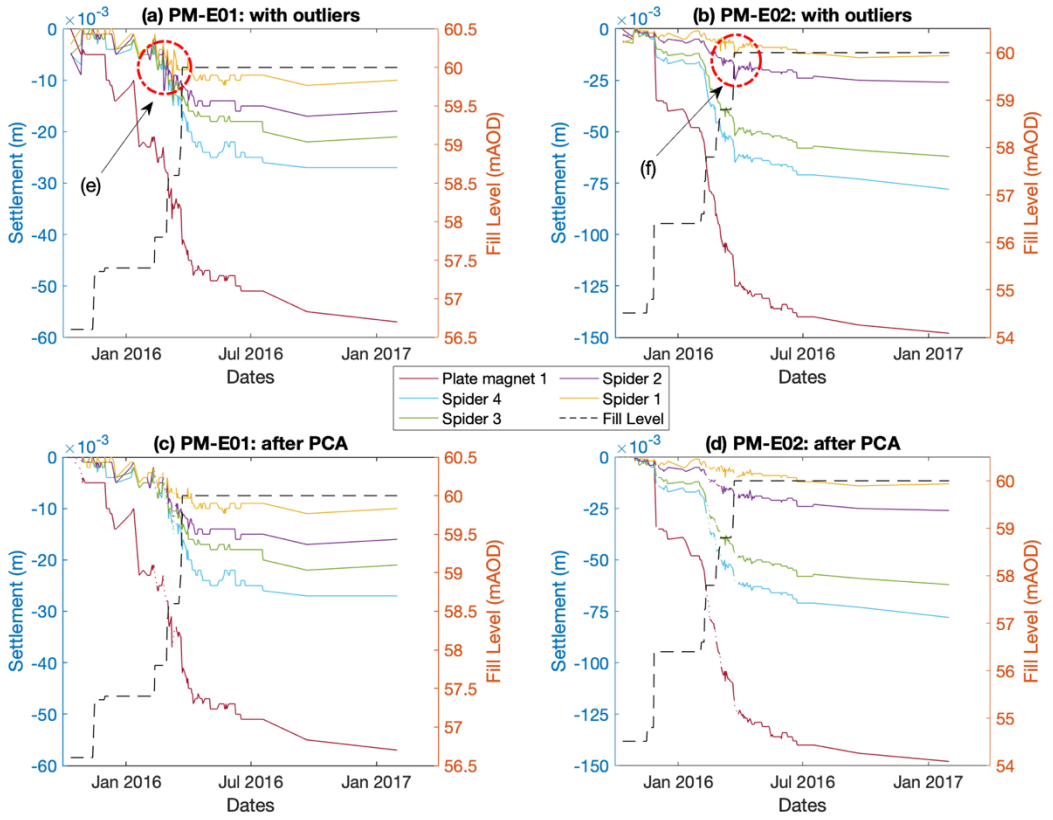
369 Hotelling T^2 metric and Q contributions can be investigated to determine the
370 contribution of all variables which cause data points to deviate outside the confidence bounds
371 [34]. Presented in Figure 5 is a Q vs T^2 influence plot for PM-E01 and PM-E02 datasets, which
372 also highlights their 95% confidence bounds. Time data points positioned outside the
373 confidence bounds, as represented by the red dotted lines, indicate that they are potential
374 outliers. Settlement values of these time data points were removed from the dataset and
375 approximated by linear interpolation. This was considered appropriate since the data sampling
376 rate was relatively high compared with the dynamics of the embankment settlement. Thus, over
377 a short time period compared to the dynamic response, the error associated with linear
378 interpolation will be small and negates the need for higher order interpolation methods. From
379 Figure 5, the first two data points recorded were outliers given their statistical deviation from
380 the confidence bounds. While it would be considered normal practice to linearly interpolate
381 between data points upon removing outliers, this was not possible given that the data points in
382 question were the first two readings taken prior to any embankment construction. Hence, a
383 value of 0 was assigned to these readings. This approach was adopted based on the time-lag
384 observed between the time at which embankment fill was placed and the corresponding
385 settlement response of the ME spiders and plate magnet. Thereafter, all remaining outliers were
386 removed, and linear interpolation was applied.
387



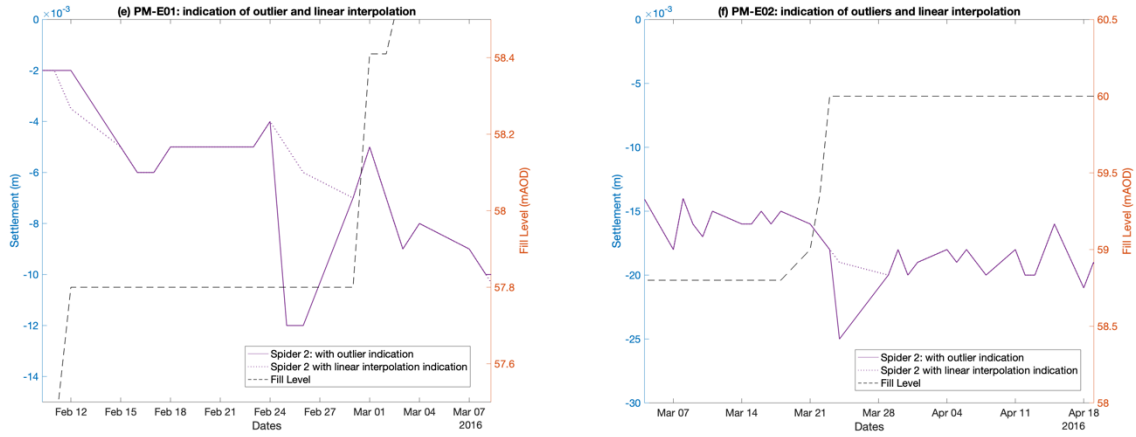
388 **Figure 5**: Influence Plot for (a) PM-E01and (b) PM-E02.

389 *5.2. Filtering and Smoothing*

390 Figure 6 shows the PM-E01 and PM-E02 datasets pre and post removal of outliers by
391 using PCA and subsequent linear interpolation, as well as highlighted areas showing outlier
392 removal details. It is apparent that there were some frequent fluctuations in the settlement
393 values, arising from factors such as measurement errors made by site operators and trafficking
394 of heavy construction machinery. In practical terms, such short-term noise within the data
395 needed to be removed by applying filtering and smoothing methods.



396



397 **Figure 6**: Settlement vs time plot - for original data (a) PM-E01 & (b) PM-E02, after outlier removal by PCA

398 (c) PM-E01 and (d) PM-E02, zoomed in indication of outliers and their removal (e) PM-E01 (f) PM-E02.

399 Following the methodology presented in section 4, given that the most appropriate

400 filtering and smoothing method was originally unknown, a range of algorithms were applied

401 to the PM-E01 and PM-E02 settlement datasets. The parameters used for each filter were

402 iteratively adjusted based on the geological knowledge and observations made by the

engineering expert monitoring the settlement data. The adjustments were made to capture the behaviour of the soil whilst removing any associated noise from the instruments. The results from applying these different filtering and smoothing methods are hereby presented.

5.2.1. Gaussian-weighted moving average

For the Gaussian-weighted moving average filtering, the window length over which data points were averaged was determined heuristically based on the original input extensometer data [46]. The value for the moving window length selected for all of the spiders and plate magnets for PM-E01 and PM-E02 were explored. It was identified that using window lengths <72 PM-E01 and <71 for PM-E02 resulted in settlement curves that were less smooth for all spiders and base magnets. However, one exception was noted for spider 1 in PM-E02, whereby a shorter window length value of 36 was required to produce a similarly smooth curve. These results (Figures 7 and 8) provide valuable insights into the settlement behaviour of the foundation soils. These findings suggest that the window length required for this filtering method is highly dependent on the nature of the raw settlement data. It is highly recommended that an iterative procedure be adopted in finding the most appropriate window length value to best suit the dataset in question.

5.2.2. Moving average smoothing

The moving average method was applied using a similar iterative approach, regarding selection of the most appropriate window length for reducing the noise content of the settlement data [47]. A shorter window length of 49 was required for all spiders and 25 for plate magnet 1 for producing smoothed data curves for PM-E01 and PM-E02, as shown in Figures 7 and 8.

5.2.3. Savitzky-Golay filtering

This method attempts to fit a polynomial curve of a specific order across a window either side of the current data point. Hence, the filter parameters were the polynomial order and the window size [48]. Trials were performed for assessing the most appropriate window length and polynomial order for the ME datasets. Based on these trial results, a polynomial order of 2 was used for both PM-E01 and PM-E02. Furthermore, window sizes of 83 and 63 were used for spiders and plate magnet 1 respectively. The resulting smoothed data curves for PM-E01 and PM-E02 are presented in Figures 7 and 8 respectively.

5.2.4. Zero-phase filtering

For zero-phase filtering, it is common practice to assume that low frequencies of the signal contain useful data, whereas high frequencies are noise [40]. Therefore, low pass filtering using the Butterworth method was adopted for cleaning the data by defining the optimum "cut-off" frequency that provides the best-fit to the original data [49]. As for previous filtering techniques, numerous trials were initially performed to define the most appropriate values for the aforementioned filter parameters. Results from these trials indicated that using low-pass filtering with cut-off frequencies of 0.015 and 0.055 for PM-E01 and PM-E02 respectively, were the most appropriate for capturing low frequency changes in the datasets, which are characteristic of soil settlement [40]. In contrast to the other filtering techniques applied in this study, the filter parameter values used in this technique were different for the two ME's. Possible reasons for this include a minor variation in soil conditions present at PM-E01 compared with PM-E02, and that slightly more engineering fill was placed over PM-E02 due to its lower initial ground level.

5.2.5. Comparison of filtering results

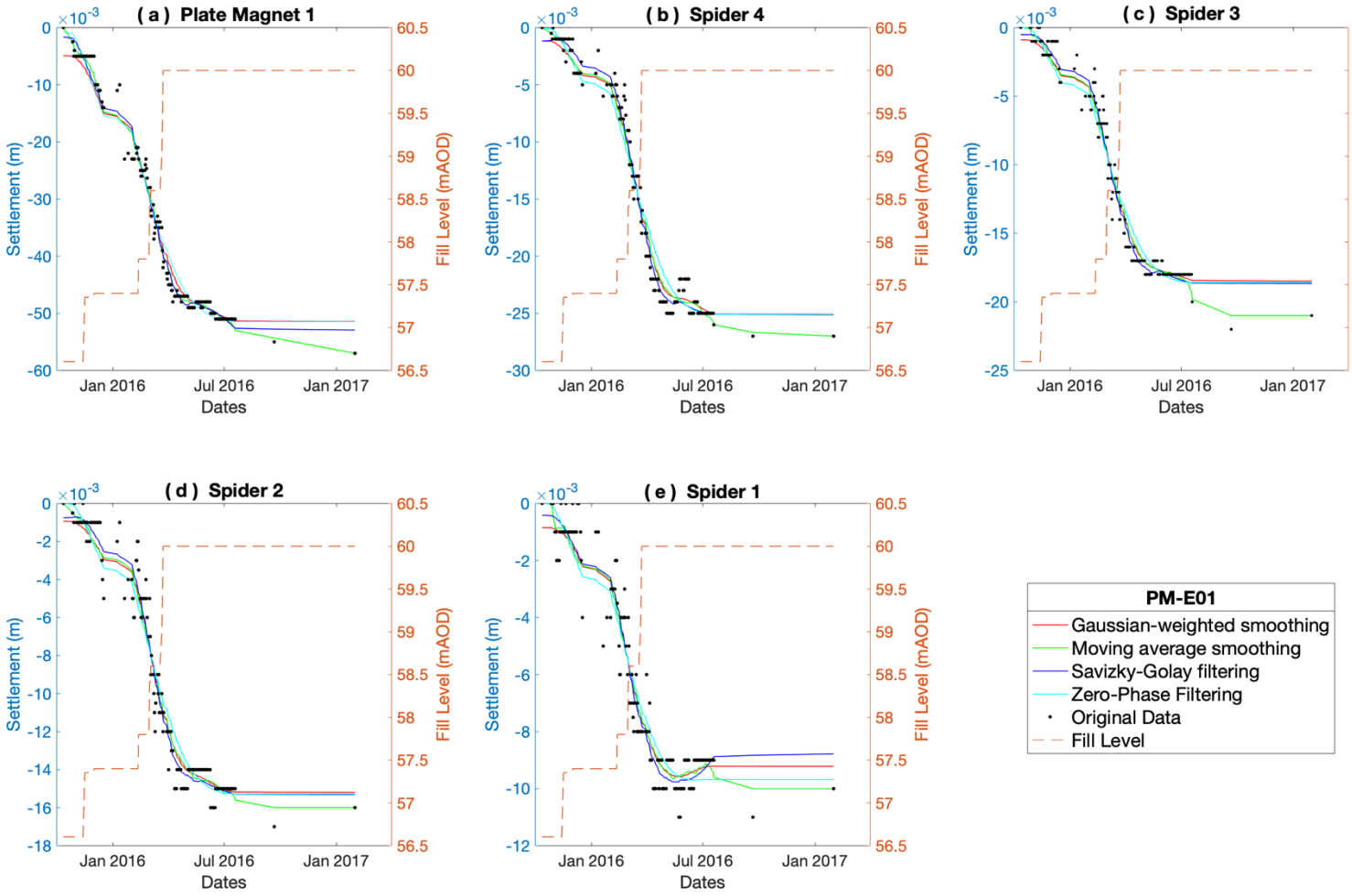
The filtering results are plotted against the original ME data in Figure 7 for PM-E01 and Figure 8 for PM-E02. It is evident that based on careful selection of values for filtering parameters, all of the filtering methods applied were successful in: 1) producing smoothed data curves that corrected for the presence of noise in the original data and 2) capturing the expected underlying trend of soil settlement at both instruments.

Based on soil consolidation theory, it is expected that for any given vertical load applied to the soil, its settlement behaviour will be characterised by a smooth exponential curve [50]. Moreover, based on soil material properties such as stiffness and permeability, there will be a maximum rate at which settlement will occur. Therefore, the variance in the raw data and filtered-smoothed data curves is due to 1) presence of noise in the data after removal of outliers and 2) resolution of the raw settlement data was 1 mm whereas filtering results showed values with a higher level of resolution. Therefore, filtering results showed a smoother trend compared with the raw data and therefore more characteristic of field behaviour.

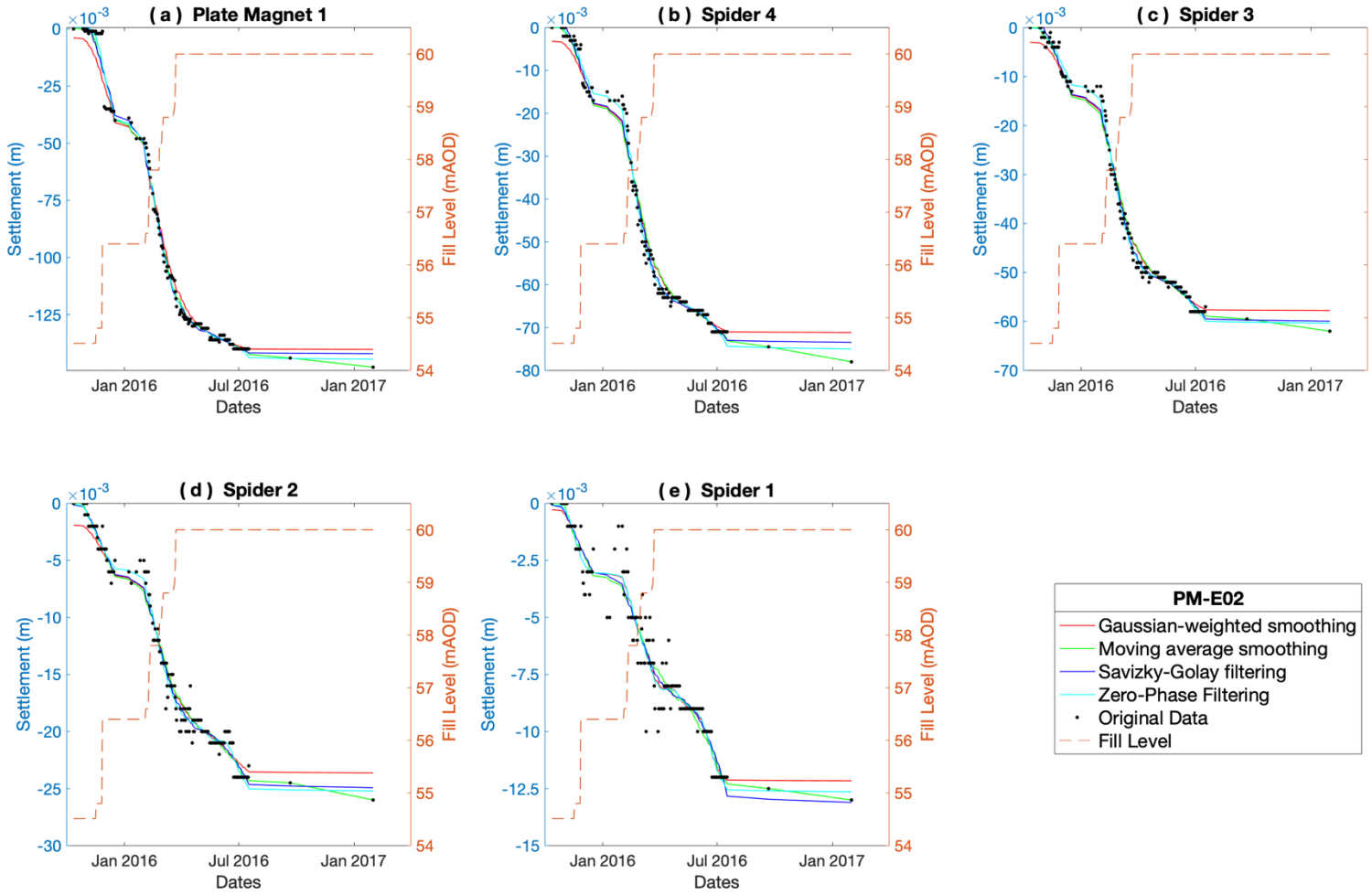
A statistical comparison of the effectiveness and reliability of these filtered results is presented in Tables 4 and 5 for PM-E01 and PM-E02, respectively. It is clear that none of the four filtering approaches used outperformed the others. These approaches were selected due to their high levels of consistency in terms of NMSE and RMSE values. Subtle variations are apparent (e.g. PM-E01 in Table 4), whereby Savitsky-Golay appeared to be slightly better. Whereas for PM-E02 (Table 5), the Zero-Phase filtering method appeared to perform better. This highlights that without implementing the four filtering approaches on the settlement data, it would not be possible to fully assess which approach is more preferable.

Based on settlement and consolidation behaviour typically expected for glacially derived lightly-overconsolidated cohesive soils, it can generally be seen that all of the filtering methods produced smoothed data curves that closely resembled the expected settlement

behaviour for Pegswood Moor embankment. Relatively high RMSE values were recorded for Plate Magnet 1 data compared with other Spiders data for all filtering and smoothing methods. However, whilst such RMSE values suggest that the filtering techniques applied were less effective, close inspection of the raw and smoothed data curves is essential as RMSE by itself is insufficient for measuring good quality filtering. Furthermore, it is worth noting that minimising the RMSE value was not the objective of this analysis, as this would indicate a perfect fit to noisy data (i.e. RMSE = 0). A similar situation was also observed for NMSE values (whereby NMSE = 1 would denote a perfect fit to noisy data) [51]. For the purpose of the settlement data presented in this study, NMSE and RMSE values were sought to reflect noise removal and also captured the underlying data trend. However, it must be emphasised that the quality of the smoothed data and associated NMSE and RMSE values highly depend on the soil conditions present, the structure being built and experienced geotechnical engineering judgement.



491 **Figure 7**: Filtering and smoothing results for PM-E01.



492

493 **Figure 8**: Filtering and smoothing results for PM-E02.

Table 4: Statistical comparison of filtering and smoothing results for PM-E01.

Statistical method//Filtering Method	**Raw Data**	**Gaussian-weighted**	**Moving-average**	**Savitzky-Golay**	**Zero-Phase**
Plate Magnet 1					
RMSE	0.0000	0.0023	0.0017	0.0020	0.0025
NMSE	1.0000	0.9833	0.9905	0.9873	0.9793
Spider 4					
RMSE	0.0000	0.0013	0.0013	0.0011	0.0017
NMSE	1.0000	0.9827	0.9816	0.9860	0.9670
Spider 3					
RMSE	0.0000	0.0010	0.0010	0.0010	0.0013
NMSE	1.0000	0.9779	0.9788	0.9812	0.9679
Spider 2					
RMSE	0.0000	0.0010	0.0010	0.0009	0.0012
NMSE	1.0000	0.9697	0.9676	0.9736	0.9564
Spider 1					
RMSE	0.0000	0.0008	0.0008	0.0008	0.0009
NMSE	1.0000	0.9463	0.9462	0.9480	0.9317

Table 5: Statistical comparison of filtering and smoothing results for PM-E02.

Statistical method//Filtering Method	**Raw Data**	**Gaussian-weighted**	**Moving-average**	**Savitzky-Golay**	**Zero-Phase**
Plate Magnet 1					
RMSE	0.0000	0.0056	0.0040	0.0044	0.0036
NMSE	1.0000	0.9879	0.9937	0.9925	0.9949
Spider 4					
RMSE	0.0000	0.0027	0.0027	0.0022	0.0017
NMSE	1.0000	0.9883	0.9889	0.9925	0.9955
Spider 3					
RMSE	0.0000	0.0021	0.0020	0.0017	0.0013
NMSE	1.0000	0.9895	0.9899	0.9934	0.9961
Spider 2					
RMSE	0.0000	0.0012	0.0011	0.0010	0.0009
NMSE	1.0000	0.9773	0.9789	0.9825	0.9865
Spider 1					
RMSE	0.0000	0.0008	0.0009	0.0009	0.0008
NMSE	1.0000	0.9526	0.9404	0.9441	0.9534

6. Discussion

Initial data exploration by PCA confirmed that soil settlement values varied with changes in the embankment fill level, as seen in Figure 4. In both ME's, the first principal component captured >95% of the total data variation, implying a strong correlation between embankment fillings and settlement. However, the amount of settlement depends on the soil

layer depth, due to effects of overburden stress and degree of consolidation. PCA was effective in identifying and removing outliers, whereby the removed data was replaced by linear interpolation. Failure to remove outlying data points from the raw dataset significantly degrades the performance of the filtering.

However, the removal of settlement outliers and data cleaning by filtering and smoothing methods is often not a straightforward process. The challenging issue with data cleaning is, in general, the risk of accidentally removing correct data points. Although this is sometimes inevitable, it can be minimised by maintaining human-in-the-loop to adjust filter parameters for an optimised result. Whilst the RMSE and NMSE metrics that have been used are useful for informing the filter performance, these are based on experienced engineering judgement. For this particular study, geotechnical knowledge of how the soils beneath the embankment were formed and their mechanical behaviour meant that it was not possible for them to experience heave during embankment loading. However, cleaning data by removing all data points that suggest heave is an act of over-conservatism, whereby there would have been a risk of falsely removing correct data.

To verify the data processing approach, it is good practice to contrast changes in correlated measurements with values that are being filtered. Hence, if changes were observed in both the filtered and correlated variable values, this questions whether such data points were outliers. For this study, given that pore water pressures within soils generally respond to compression associated with embankment construction, comparisons were also made with VWP datasets (up to 1st June 2016) collected for Pegswood Moor embankment.

Figures 9 and 10 present VWP data collected for PM-P02 and PM-P01 respectively, whereby comparisons in pore water pressure were made with settlement measurements recorded for spiders at corresponding depths within PM-E01 and PM-E02 respectively. It should be noted that the pore water pressures recorded by both PM-P01 and PM-P02 were not

corrected with respect to the effect of the instruments settling with time, due to embankment construction and that the tips of these instruments were within compressible strata rather than rock or dense soils (as for PM-E01 and PM-E02).

In Figures 9 and 10, it can generally be seen that during the earlier stages of embankment construction, pore water pressures increased at approximately the same time as individual fill levels being applied and the corresponding soil settlement response. For VWP sensor PM-P02B, pore water pressure measurements were largely recorded to be negative – indicating suctions. This suggests that the soil was partially saturated and that the local groundwater level was located at a depth lower than that for this particular sensor.

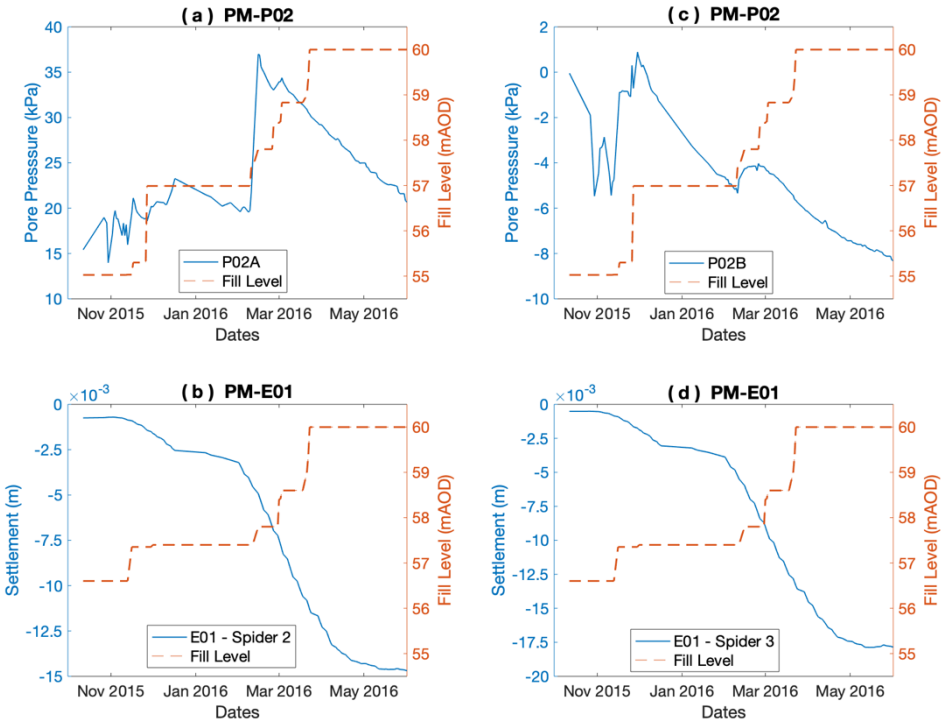


Figure 9: Combined Piezometer and Extensometer Data for (a and c) PM-P02 and (b and d) PM-E01.

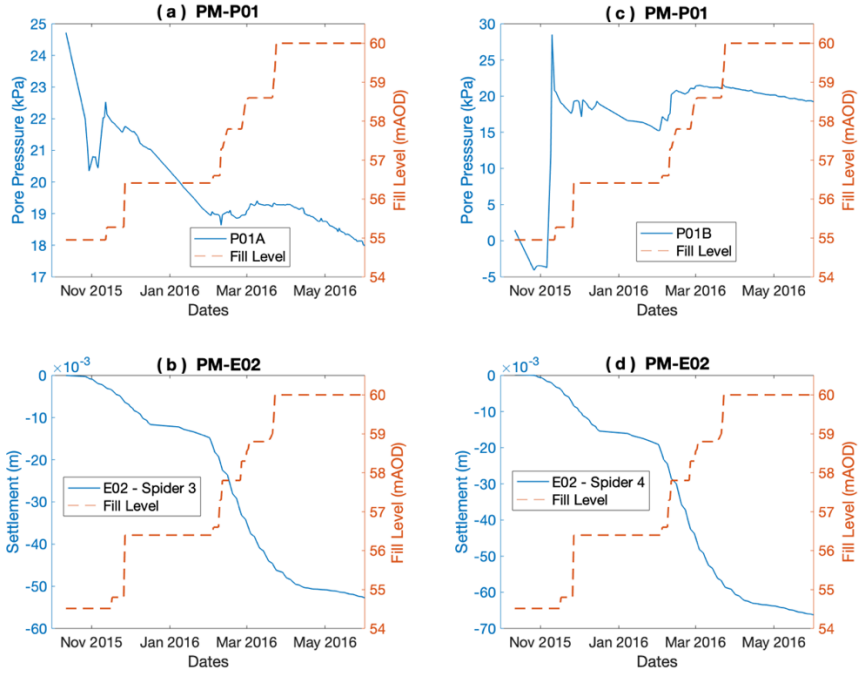


Figure 10: Combined Piezometer and Extensometer Data for (a and c) PM-P01 and (b and d) PM-E02.

For some stages of embankment filling, there was not a clear immediate settlement response. Hence, it has proved useful to use the VWP data for cross-checking the timings of embankment fill layers being placed. However, in the later stages of embankment construction, sharp peaks in pore water pressures became less significant. This can be explained by the presence of the pre-fabricated vertical drains, which rapidly increased the rate of soil consolidation and dissipation of excess pore water pressures. Also, the effects of construction compaction and trafficking of site machinery further contributed towards the consolidation of the embankment and suppression of pore water pressures. Relatively few time lags were observed between the application of embankment fill and settlement – pore water pressure responses, which were likely due to the effects of the soil's low hydraulic conductivities. Whilst the comparison of VWP and settlement data is following good practice, the resolution of the VWP dataset alone was insufficient for identifying outliers in the settlement data. However, the VWP dataset was useful in confirming the timings of the settlement response.

Although the ME and VWP instrumentation were only in operation during the construction of the Pegswood Moor embankment, the raw data and filtered-smoothed data curves suggest that primary consolidation and a large proportion of secondary consolidation had completed after the final embankment fill increment and during the final hold period. Encouragingly, the data pre-processing results and smoothed settlement data curves produced for this case study appear to be of similarly good quality as those produced by Kelly et al. [5] for a purpose built embankment with a much denser VWP and ME sensor network.

7. Conclusions and Recommendations

This paper highlights that data pre-processing is a vital activity for interpreting real-time settlement data collected for earth embankments. Outliers and noise in the data mask the underlying soil settlement behaviour and can ultimately lead to less informed decisions on site during the embankment construction (such as duration of hold periods between filling). In addition, outliers and noise in the datasets can lead to unreliable data-driven models and their prediction capabilities.

Simple pre-processing methods including PCA provided a useful indicator of data outliers and noise in the embankment settlement. Results from commonly used cleaning, filtering and smoothing algorithms demonstrated that they well captured the overall expected and observed settlement behaviour of the Pegswood Moor embankment. However, it should be noted that this study only used simple smoothing based algorithms for data cleaning. This was motivated by the primary reasons of simplicity and usability. If these simple algorithms are not suitable for other settlement datasets, more complex cleaning algorithms are available, but it would be more challenging to establish filtering parameters. Whilst the smoothing algorithms are transferable to processing settlement data for other earth embankments, the

values assigned to filtering parameters for this study are not due to differences in soil conditions and instrumentation installation details.

Due to natural spatial variations in ground conditions, it is common practice to collect as much data as possible from individual monitoring instruments and to invest in installing more of them – especially if the data generated is difficult to interpret. However, if interpretation can be improved this may assist in reducing and thereby optimising the number of instruments required. Based on the findings from this study, comparatively small-sized datasets generated from geotechnical site-based monitoring instruments provides an effective indication of embankment settlement. This negates the need to install extensive arrays of sophisticated instrumentation across embankment structures and is a cost-effective approach for data-driven modelling.

It is strongly recommended that professional geotechnical engineers spend sufficient time in understanding raw settlement data and adopt pre-processing and filtering methods for enhancing data quality. This provides more detailed insights into subtle variations within embankment settlement data. In general, data pre-processing is an initial step for data-driven modelling to resolve issues such as outliers and missing data. Without this, data outliers can obscure the true data trends. Therefore, the data analysis performed in this paper was also beneficial in providing site engineers with improved short-term information concerning embankment settlement dynamics. After pre-processing, the data can be used to implement pattern recognition and machine learning concepts to evaluate the change in the condition of the infrastructure over time. The predictive capabilities of a machine learning model can then be utilised to evaluate the life expectancy of the infrastructure. If no data pre-processing is undertaken, the quality of any subsequent machine learning or predictive modelling will be degraded.

Acknowledgements

Thanks go to Teesside University for funding this research. In addition, the authors would like to thank Northumberland County Council and AECOM Environment and Ground Engineering (Newcastle upon Tyne, UK) for providing all of the monitoring data from the Morpeth Northern Bypass.

References

[1] F. Din-Houn Lau, L.J. Butler, N.M. Adams, M.Z.E.B. Elshafie, M.A. Girolami, Real-time statistical modelling of data generated from self-sensing bridges, Proc. Inst. Civ. Eng. - Smart Infrastruct. Constr. 171 (2018) 3–13. https://doi.org/10.1680/jsmic.17.00023.

[2] A.H. Alavi, A.H. Gandomi, Big data in civil engineering, Elsevier, 2017. https://www.sciencedirect.com/science/article/pii/S0926580516305246 (accessed February 14, 2019).

[3] E.J. Oughton, W. Usher, P. Tyler, J.W. Hall, Infrastructure as a Complex Adaptive System, Complexity. 2018 (2018) 3427826. https://doi.org/10.1155/2018/3427826.

[4] A.J. Hargreaves, M. Cavada, C.D.F. Rogers, Briefing: Engineering for the far future: rethinking the value proposition, Proc. Inst. Civ. Eng. - Eng. Sustain. 173 (2020) 3–7. https://doi.org/10.1680/jensu.19.00020.

[5] R.B. Kelly, S.W. Sloan, J.A. Pineda, G. Kouretzis, J. Huang, Outcomes of the Newcastle symposium for the prediction of embankment behaviour on soft soil, Comput. Geotech. 93 (2018) 9–41. https://doi.org/10.1016/j.compgeo.2017.08.005.

[6] K.F. Chan, B.M. Poon, D. Perera, Prediction of embankment performance using numerical analyses – Practitioner's approach, Comput. Geotech. 93 (2018) 163–177. https://doi.org/10.1016/j.compgeo.2017.07.012.

[7] L. Ho, B. Fatahi, One-dimensional consolidation analysis of unsaturated soils subjected to time-dependent loading, Int. J. Geomech. 16 (2016) 04015052. https://doi.org/10.1061/(ASCE)GM.1943-5622.0000504.

[8] D. Zheng, J. Huang, D.Q. Li, R. Kelly, S.W. Sloan, Embankment prediction using testing data and monitored behaviour: A Bayesian updating approach, Comput. Geotech. 93 (2018) 150–162. https://doi.org/10.1016/j.compgeo.2017.05.003.

[9] N. Müthing, C. Zhao, R. Hölter, T. Schanz, Settlement prediction for an embankment on soft clay, Comput. Geotech. 93 (2018) 87–103. https://doi.org/10.1016/j.compgeo.2017.06.002.

[10] R. Kelly, J. Huang, Bayesian updating for one-dimensional consolidation measurements, Can. Geotech. J. 52 (2015) 1318–1330. https://doi.org/10.1139/cgj-2014-0338.

[11] C.R. Farrar, K. Worden, Structural Health Monitoring: A Machine Learning Perspective, Chichester, UK, 2012. https://doi.org/10.1002/9781118443118.

[12] E. Figueiredo, A. Santos, Machine Learning Algorithms for Damage Detection, in: 2018: pp. 1–39. https://doi.org/10.1142/9781786344977_0001.

[13] J. Dunnicliff, W.A. Marr, J. Standing, Chapter 94 Principles of geotechnical monitoring, in: ICE Man. Geotech. Eng. Vol. II, n.d.: pp. 1363–1377. https://doi.org/10.1680/moge.57098.1363.

[14] C. Cernuda, On the relevance of preprocessing in predictive maintenance for dynamic systems, in: Predict. Maint. Dyn. Syst. Adv. Methods, Decis. Support Tools Real-World Appl., Springer International Publishing, 2019: pp. 53–92. https://doi.org/10.1007/978-3-030-05645-2_3.

[15] X. Chu, I.F. Ilyas, S. Krishnan, J. Wang, Data cleaning: Overview and emerging challenges, in: Proc. ACM SIGMOD Int. Conf. Manag. Data, Association for

Computing Machinery, 2016: pp. 2201–2206. https://doi.org/10.1145/2882903.2912574.

[16] A. Klein, W. Lehner, Representing data quality in sensor data streaming environments, J. Data Inf. Qual. 1 (2009) 1–28. https://doi.org/10.1145/1577840.1577845.

[17] M. Mezzanzanica, R. Boselli, M. Cesarini, F. Mercorio, A model-based evaluation of data quality activities in KDD, Inf. Process. Manag. 51 (2015) 144–166. https://doi.org/10.1016/j.ipm.2014.07.007.

[18] S. Krishnan, D. Haas, M.J. Franklin, E. Wu, Towards reliable interactive data cleaning: A user survey and recommendations, in: HILDA 2016 - Proc. Work. Human-In-the-Loop Data Anal., Association for Computing Machinery, Inc, New York, New York, USA, 2016: pp. 1–5. https://doi.org/10.1145/2939502.2939511.

[19] Y. Gong, Y.H. Chok, Predicted and measured behaviour of a test embankment on Ballina clay, Comput. Geotech. 93 (2018) 178–190. https://doi.org/10.1016/j.compgeo.2017.06.003.

[20] S. Fallah, Application of Machine Learning in Geotechnics, 2018. https://www.researchgate.net/publication/327338485_Application_of_Machine_Learning_in_Geotechnics.

[21] Y. Cai, Y. Chen, Z. Cao, C. Ren, A combined method to predict the long-term settlements of roads on soft soil under cyclic traffic loadings, Acta Geotech. 13 (2018) 1215–1226. https://doi.org/10.1007/s11440-017-0616-3.

[22] Digimap, (n.d.). https://digimap.edina.ac.uk/ (accessed April 8, 2020).

[23] J. Dunnicliff, Chapter 95 Types of geotechnical instrumentation and their usage, in: ICE Man. Geotech. Eng. Vol. II, n.d.: pp. 1379–1403. https://doi.org/10.1680/moge.57098.1379.

[24] Magnetic Probe Extensometer - Soil Instruments, (n.d.).

https://soilinstruments.com/products/extensometers/magnetic-probe-extensometer/ (accessed March 10, 2020).

[25] W9 Vibrating Wire Piezometer - Soil Instruments, (n.d.). https://soilinstruments.com/products/water-monitors-piezometers-meters/vibrating-wire-piezometer/ (accessed March 10, 2020).

[26] S.A. Saputro, A. Setyo Muntohar, H. Jiun Liao, Ground settlement prediction of embankment treated with prefabricated vertical drains in soft soil, in: MATEC Web Conf., EDP Sciences, 2018. https://doi.org/10.1051/matecconf/201819503014.

[27] A. Asaoka, Observational procedure of settlement prediction., Soils Found. 18 (1978) 87–101. https://doi.org/10.3208/sandf1972.18.4_87.

[28] Y. Hu, H. Chen, G. Li, H. Li, R. Xu, J. Li, A statistical training data cleaning strategy for the PCA-based chiller sensor fault detection, diagnosis and data reconstruction method, Energy Build. 112 (2016) 270–278. https://doi.org/10.1016/j.enbuild.2015.11.066.

[29] M. Li, Y. Shen, Q. Ren, H. Li, A new distributed time series evolution prediction model for dam deformation based on constituent elements, Adv. Eng. Informatics. 39 (2019) 41–52. https://doi.org/10.1016/j.aei.2018.11.006.

[30] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey, J. Netw. Comput. Appl. 73 (2016) 57–81. https://doi.org/10.1016/J.JNCA.2016.08.002.

[31] X. Wang, C. Wang, Time Series Data Cleaning: A Survey, IEEE Access. 8 (2020) 1866–1881. https://doi.org/10.1109/ACCESS.2019.2962152.

[32] D. Sen, A. Aghazadeh, A. Mousavi, S. Nagarajaiah, R. Baraniuk, A. Dabak, Data-driven semi-supervised and supervised learning algorithms for health monitoring of pipes, Mech. Syst. Signal Process. 131 (2019) 524–537.

https://doi.org/10.1016/j.ymssp.2019.06.003.

[33] L. Mujica, J. Rodellar, A. Fernández, A. Gü emes, Q-statistic and T 2-statistic PCA-based measures for damage assessment in structures, (n.d.). https://doi.org/10.1177/1475921710388972.

[34] D. Ballabio, A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure, Chemom. Intell. Lab. Syst. 149 (2015) 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003.

[35] J. Vitola, F. Pozo, D. Tibaduiza, M. Anaya, A Sensor Data Fusion System Based on k-Nearest Neighbor Pattern Classification for Structural Health Monitoring Applications, Sensors. 17 (2017) 417. https://doi.org/10.3390/s17020417.

[36] I.T. Jollife, J. Cadima, Principal component analysis: A review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016). https://doi.org/10.1098/rsta.2015.0202.

[37] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[38] G.F. Sirca, H. Adeli, System identification in structural engineering, Sci. Iran. 19 (2012) 1355–1364. https://doi.org/10.1016/j.scient.2012.09.002.

[39] Y. Ying, J.H. Garrett, I.J. Oppenheim, L. Soibelman, J.B. Harley, J. Shi, Y. Jin, Toward data-driven structural health monitoring: Application of machine learning and signal processing to damage detection, in: J. Comput. Civ. Eng., 2013: pp. 667–680. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000258.

[40] A. Swider, E. Pedersen, Comparison of delayless digital filtering algorithms and their application to multi-sensor signal processing, Trans. Inst. Meas. Control. 41 (2019) 2338–2351. https://doi.org/10.1177/0142331218799148.

[41] J.-L. Liu, J.-Y. Zheng, X.-J. Wei, F.-Y. Liao, Y.-P. Luo, A new instantaneous

frequency extraction method for nonstationary response signals in civil engineering structures, J. Low Freq. Noise, Vib. Act. Control. 37 (2018) 834–848. https://doi.org/10.1177/1461348418790534.

[42] P. Janert, Data Analysis with Open Source Tools, O'Reilly Media, Incorporated, 2010. http://books.google.co.kr/books?id=mTXnXCLXJYgC&printsec=frontcover&dq=data+analysis+with+open+source+tools&hl=&cd=1&source=gbs_api (accessed March 11, 2020).

[43] M. de Oliveira, N. Araujo, R. da Silva, T. da Silva, J. Epaarachchi, Use of Savitzky–Golay Filter for Performances Improvement of SHM Systems Based on Neural Networks and Distributed PZT Sensors, Sensors. 18 (2018) 152. https://doi.org/10.3390/s18010152.

[44] J.M. Giron-Sierra, Digital Signal Processing with Matlab Examples, Volume 1, Springer Singapore, Singapore, 2017. https://doi.org/10.1007/978-981-10-2534-1.

[45] Mathworks, MATLAB version 9.7.0.1319299 (R2019b) Update 5, Massachusetts, U.S.A. (2019).

[46] Smooth noisy data - MATLAB smoothdata - MathWorks United Kingdom, (n.d.). https://uk.mathworks.com/help/matlab/ref/smoothdata.html#bvhejau-method%5D (accessed March 12, 2020).

[47] Smooth response data - MATLAB smooth - MathWorks United Kingdom, (n.d.). https://uk.mathworks.com/help/curvefit/smooth.html#mw_ad6b65fd-4dac-46c4-a649-a7a0b301eb80 (accessed March 12, 2020).

[48] Savitzky-Golay filtering - MATLAB sgolayfilt - MathWorks United Kingdom, (n.d.). https://uk.mathworks.com/help/signal/ref/sgolayfilt.html (accessed April 29, 2020).

[49] Zero-phase digital filtering - MATLAB filtfilt - MathWorks United Kingdom, (n.d.).

https://uk.mathworks.com/help/signal/ref/filtfilt.html (accessed March 12, 2020).

[50] I. Smith, Smith's elements of soil mechanics, 9th ed., Wiley-Blackwell, Chichester, 2014.

[51] MATLAB goodnessOfFit - MathWorks United Kingdom, (n.d.). https://uk.mathworks.com/help/ident/ref/goodnessoffit.html (accessed March 31, 2020).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: